PATENT COOPERATION TREATY

From the INTERNATIONAL SEARCHING AUTHORITY

PCT

NOTIFICATION OF TRANSMITTAL OF THE INTERNATIONAL SEARCH REPORT OR THE DECLARATION

(PCT Rule 44.1)

To:
INTERNATIONAL BUSINESS MACHINES CORP. - IBM
Attn. KLETT, Peter M.
Säumerstraße 4
8803 Rüschlikon
SWITZERLAND

Date of mailing *(day/month/year)* 17/12/1997	
Applicant's or agent's file reference SZ9-97-003	**FOR FURTHER ACTION** See paragraphs 1 and 4 below
International application No. PCT/IB 97/00394	International filing date *(day/month/year)* 10/04/1997

Applicant

INTERNATIONAL BUSINESS MACHINES CORPORATION et al.

1. [X] The applicant is hereby notified that the International Search Report has been established and is transmitted herewith.

Filing of amendments and statement under Article 19
The applicant is entitled, if he so wishes, to amend the claims of the International Application (see Rule 46):

When? The time limit for filing such amendments is normally 2 months from the date of transmittal of the International Search Report; however, for more details, see the notes on the accompanying sheet.

Where? Directly to the International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland
Fascimile No.: (41-22) 740.14.35

For more detailed instructions, see the notes on the accompanying sheet.

2. [ ] The applicant is hereby notified that no International Search Report will be established and that the declaration under Article 17(2)(a) to that effect is transmitted herewith.

3. [ ] **With regard to the protest** against payment of (an) additional fee(s) under Rule 40.2, the applicant is notified that:

[ ] the protest together with the decision thereon has been transmitted to the International Bureau together with the applicants's request to forward the texts of both the protest and the decision thereon to the designated Offices.

[ ] no decision has been made yet on the protest; the applicant will be notified as soon as a decision is made.

4. **Further action(s):** The applicant is reminded of the following:

Shortly after **18 months** from the priority date, the international application will be published by the International Bureau. If the applicant wishes to avoid or postpone publication, a notice of withdrawal of the international application, or of the priority claim, must reach the International Bureau as provided in Rules 90*bis*.1 and 90*bis*.3, respectively, before the completion of the technical preparations for international publication.

Within **19 months** from the priority date, a demand for international preliminary examination must be filed if the applicant wishes to postpone the entry into the national phase until 30 months from the priority date (in some Offices even later).

Within **20 months** from the priority date, the applicant must perform the prescribed acts for entry into the national phase before all designated Offices which have not been elected in the demand or in a later election within 19 months from the priority date or could not be elected because they are not bound by Chapter II.

Name and mailing address of the International Searching Authority European Patent Office, P.B. 5818 Patentlaan 2 NL-2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Lucia Van Pinxteren

Form PCT/ISA/220 (January 1994)

NOTES TO FORM PCT/ISA/220

These Notes are intended to give the basic instructions concerning the filing of amendments under article 19. The Notes are based on the requirements of the Patent Cooperation Treaty, the Regulations and the Administrative Instructions under that Treaty. In case of discrepancy between these Notes and those requirements, the latter are applicable. For more detailed information, see also the PCT Applicant's Guide, a publication of WIPO.

In these Notes, "Article", "Rule", and "Section" refer to the provisions of the PCT, the PCT Regulations and the PCT Administrative Instructions respectively.

INSTRUCTIONS CONCERNING AMENDMENTS UNDER ARTICLE 19

The applicant has, after having received the international search report, one opportunity to amend the claims of the international application. It should however be emphasized that, since all parts of the international application (claims, description and drawings) may be amended during the international preliminary examination procedure, there is usually no need to file amendments of the claims under Article 19 except where, e.g. the applicant wants the latter to be published for the purposes of provisional protection or has another reason for amending the claims before international publication. Furthermore, it should be emphasized that provisional protection is available in some States only.

What parts of the International application may be amended?

Under Article 19, only the claims may be amended.

During the international phase, the claims may also be amended (or further amended) under Article 34 before the International Preliminary Examining Authority. The description and drawings may only be amended under Article 34 before the International Examining Authority.

Upon entry into the national phase, all parts of the international application may be amended under Article 28 or, where applicable, Article 41.

When? Within 2 months from the date of transmittal of the international search report or 16 months from the priority date, whichever time limit expires later. It should be noted, however, that the amendments will be considered as having been received on time if they are received by the International Bureau after the expiration of the applicable time limit but before the completion of the technical preparations for international publication (Rule 46.1).

Where not to file the amendments?

The amendments may only be filed with the International Bureau and not with the receiving Office or the International Searching Authority (Rule 46.2).

Where a demand for international preliminary examination has been/is filed, see below.

How? Either by cancelling one or more entire claims, by adding one or more new claims or by amending the text of one or more of the claims as filed.

A replacement sheet must be submitted for each sheet of the claims which, on account of an amendment or amendments, differs from the sheet originally filed.

All the claims appearing on a replacement sheet must be numbered in Arabic numerals. Where a claim is cancelled, no renumbering of the other claims is required. In all cases where claims are renumbered, they must be renumbered consecutively (Administrative Instructions, Section 205(b)).

The amendments must be made in the language in which the international application is to be published.

What documents must/may accompany the amendments?

Letter (Section 205(b)):

The amendments must be submitted with a letter.

The letter will not be published with the international application and the amended claims. It should not be confused with the "Statement under Article 19(1)" (see below, under "Statement under Article 19(1)").

The letter must be in English or French, at the choice of the applicant. However, if the language of the international application is English, the letter must be in English; if the language of the international application is French, the letter must be in French.

The letter must indicate the differences between the claims as filed and the claims as amended. It must, in particular, indicate, in connection with each claim appearing in the international application (it being understood that identical indications concerning several claims may be grouped), whether

(i) the claim is unchanged;

(ii) the claim is cancelled;

(iii) the claim is new;

(iv) the claim replaces one or more claims as filed;

(v) the claim is the result of the division of a claim as filed.

The following examples illustrate the manner in which amendments must be explained in the accompanying letter:

1. [Where originally there were 48 claims and after amendment of some claims there are 51]:
"Claims 1 to 29, 31, 32, 34, 35, 37 to 48 replaced by amended claims bearing the same numbers; claims 30, 33 and 36 unchanged; new claims 49 to 51 added."

2. [Where originally there were 15 claims and after amendment of all claims there are 11]:
"Claims 1 to 15 replaced by amended claims 1 to 11."

3. [Where originally there were 14 claims and the amendments consist in cancelling some claims and in adding new claims]:
"Claims 1 to 6 and 14 unchanged; claims 7 to 13 cancelled; new claims 15, 16 and 17 added." or
"Claims 7 to 13 cancelled; new claims 15, 16 and 17 added; all other claims unchanged."

4. [Where various kinds of amendments are made]:
"Claims 1-10 unchanged; claims 11 to 13, 18 and 19 cancelled; claims 14, 15 and 16 replaced by amended claim 14; claim 17 subdivided into amended claims 15, 16 and 17; new claims 20 and 21 added."

"Statement under article 19(1)" (Rule 46.4)

The amendments may be accompanied by a statement explaining the amendments and indicating any impact that such amendments might have on the description and the drawings (which cannot be amended under Article 19(1)).

The statement will be published with the international application and the amended claims.

It must be in the language in which the international appplication is to be published.

It must be brief, not exceeding 500 words if in English or if translated into English.

It should not be confused with and does not replace the letter indicating the differences between the claims as filed and as amended. It must be filed on a separate sheet and must be identified as such by a heading, preferably by using the words "Statement under Article 19(1)."

It may not contain any disparaging comments on the international search report or the relevance of citations contained in that report. Reference to citations, relevant to a given claim, contained in the international search report may be made only in connection with an amendment of that claim.

Consequence if a demand for international preliminary examination has already been filed

If, at the time of filing any amendments under Article 19, a demand for international preliminary examination has already been submitted, the applicant must preferably, at the same time of filing the amendments with the International Bureau, also file a copy of such amendments with the International Preliminary Examining Authority (see Rule 62.2(a), first sentence).

Consequence with regard to translation of the international application for entry into the national phase

The applicant's attention is drawn to the fact that, where upon entry into the national phase, a translation of the claims as amended under Article 19 may have to be furnished to the designated/elected Offices, instead of, or in addition to, the translation of the claims as filed.

For further details on the requirements of each designated/elected Office, see Volume II of the PCT Applicant's Guide.

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference	**FOR FURTHER ACTION** see Notification of Transmittal of International Search Report (Form PCT/ISA/220) as well as, where applicable, item 5 below.	
SZ9-97-003		
International application No.	International filing date *(day/month/year)*	(Earliest) Priority Date *(day/month/year)*
PCT/IB 97/ 00394	10/04/1997	

Applicant

INTERNATIONAL BUSINESS MACHINES CORPORATION et al.

This International Search Report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This International Search Report consists of a total of 3 sheets.

[X] It is also accompanied by a copy of each prior art document cited in this report.

1. [ ] **Certain claims were found unsearchable** (see Box I).

2. [ ] **Unity of invention is lacking** (see Box II).

3. [ ] The international application contains disclosure of a **nucleotide and/or amino acid sequence listing** and the international search was carried out on the basis of the sequence listing
   - [ ] filed with the international application.
   - [ ] furnished by the applicant separately from the international application,
     - [ ] but not accompanied by a statement to the effect that it did not include matter going beyond the disclosure in the international application as filed.
   - [ ] Transcribed by this Authority

4. With regard to the **title,**
   - [ ] the text is approved as submitted by the applicant
   - [X] the text has been established by this Authority to read as follows:

   METHOD AND MEANS FOR DETERMINING THE USED BANDWIDTH ON A CONNECTION ✓

5. With regard to the **abstract,**
   - [X] the text is approved as submitted by the applicant
   - [ ] the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box III. The applicant may, within one month from the date of mailing of this International Search Report, submit comments to this Authority.

6. The figure of the **drawings** to be published with the abstract is:

   Figure No. 1
   - [X] as suggested by the applicant.
   - [ ] because the applicant failed to suggest a figure.
   - [ ] because this figure better characterizes the invention.
   - [ ] None of the figures.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/IB 97/00394

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04Q11/04 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04Q H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SHIMOKOSHI K: "EVALUATION OF POLICING MECHANISMS FOR ATM NETWORKS" IEICE TRANSACTIONS ON COMMUNICATIONS, vol. E76-B, no. 11, 1 November 1993, pages 1341-1351, XP000425067 see page 1345, left-hand column, line 10 - page 1346, right-hand column, line 34 --- -/--	1,12

D1

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
2 December 1997	17/12/1997

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Veen, G

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	KIYOSHI SHIMOKOSHI: "PERFORMANCE COMPARISON OF BANDWIDTH ALLOCATION MECHANISMS FOR LAN/MAN INTERWORKING THROUGH AN ATM NETWORK" SERVING HUMANITY THROUGH COMMUNICATIONS. SUPERCOM/ICC, NEW ORLEANS, MAY 1 - 5, 1994, vol. VOL. 3, no. -, 1 May 1994, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, pages 1405-1411, XP000438728 see page 1406, left-hand column, line 40 - right-hand column, line 30 ---	1,12
A	GONG F ET AL: "STUDY OF A TWO-LEVEL FLOW CONTROL SCHEME AND BUFFERING STRATEGIES" PROCEEDINGS OF THE CONFERENCE ON COMPUTER COMMUNICATIONS (INFOCOM), TORONTO, JUNE 12 - 16, 1994, vol. 3, 12 June 1994, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, pages 1224-1233, XP000496585 see page 1226, right-hand column, line 10 - line 55 -----	1,12

INFORMATION ON TIME LIMITS FOR ENTERING THE NATIONAL PHASE

The applicant is reminded that the "national phase" must be entered before each of the designated Offices indicated in the Notification of Receipt of Record Copy (Form PCT/IB/301) by paying national fees and furnishing translations, as prescribed by the applicable national laws.

The time limit for performing these procedural acts is **20 MONTHS** from the priority date or, for those designated States which the applicant elects in a demand for international preliminary examination or in a later election, **30 MONTHS** from the priority date, provided that the election is made before the expiry of 19 months from the priority date. Some designated (or elected) Offices have fixed time limits which expire even later than 20 or 30 months from the priority date. In other Offices an extension of time or grace period, in some cases upon payment of an additional fee, is available.

In addition to these procedural acts, the applicant may also have to comply with other special requirements applicable in certain Offices. **It is the applicant's responsibility** to ensure that the necessary steps to enter the national phase are taken in a timely fashion. Most designated Offices do not issue reminders to applicants in connection with the entry into the national phase.

For detailed information about the procedural acts to be performed to enter the national phase before each designated Office, the applicable time limits and possible extensions of time or grace periods, and any other requirements, see the relevant Chapters of Volume II of the PCT Applicant's Guide. Information about the requirements for filing a demand for international preliminary examination is set out in Chapter IX of Volume I of the PCT Applicant's Guide.

Note that since ES is not bound by PCT Chapter II (which provides for the international preliminary examination procedure), that State cannot be elected in a demand for international preliminary examination. In the case of the designation of ES for a national patent, the applicant must thus always enter the national phase before the national Office of that State before the expiration of 20 months from the priority date. In the case of the designation of ES for a European patent, however, the 31-month time limit applies in respect of that designation if at least one other State designated for a European patent is also elected within the 19-month period.*

Note also that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

---

* CH and LI became bound by PCT Chapter II on 1 September 1995. GR became bound by PCT Chapter II on 7 September 1996. Therefore, CH and LI may be elected in a demand or a later election filed on or after 1 September 1995, and GR may be elected in a demand or a later election filed on or after 7 September 1996, regardless of the filing date of the international application. (See 2nd paragraph above.)

CONFIRMATION OF PRECAUTIONARY DESIGNATIONS

This notification lists only specific designations made under Rule 4.9(a) in the request. It is important to check that these designations are correct. Errors in designations can be corrected where precautionary designations have been made under Rule 4.9(b). The applicant is hereby reminded that any precautionary designations may be confirmed according to Rule 4.9(c) before the expiration of 15 months from the priority date. If it is not confirmed, it will automatically be regarded as withdrawn by the applicant. There will be no reminder and no invitation. Confirmation of a designation consists of the filing of a notice specifying the designated State concerned (with an indication of the kind of protection or treatment desired) and the payment of the designation and confirmation fees. Confirmation must reach the receiving Office within the 15-month time limit.

REQUIREMENTS REGARDING PRIORITY DOCUMENTS

For applicants who have not yet complied with the requirements regarding priority documents the following is recalled.

Where the priority of an earlier national (i.e., national or regional) application is claimed, the applicant must submit a copy of the said national application, certified by the authority with which it was filed ("the priority document") to the receiving Office (which will transmit it to the International Bureau) or directly to the International Bureau, before the expiration of 16 months from the priority date (Rule 17.1).

Where the priority document is issued by the receiving Office, the applicant may, instead of submitting the priority document, request the receiving Office to prepare and transmit the priority document to the International Bureau. Such a request must be made before the expiration of the 16-month time limit.

It is recalled that, where several priorities are claimed, the priority date to be considered for the purposes of computing the 16-month time limit is the filing date of the earliest application whose priority is claimed.

If the priority document concerned is not submitted to the International Bureau before the expiration of the 16-month time limit, or if the request to the receiving Office to transmit the priority document has not been made (and the corresponding fee, if any, paid) before the expiration of this time limit, any designated State may disregard the priority claim.

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF RECEIPT OF RECORD COPY

(PCT Rule 24.2(a))

ZURICH-PATENT

23 IV 97 8:45

To:

KLETT, Peter, Michael
International Business Machines Corporation
Saeumerstrasse 4
CH-8803 Rueschlikon
SUISSE

Date of mailing (day/month/year)	IMPORTANT NOTIFICATION
16 April 1997 (16.04.97)	
Applicant's or agent's file reference	**International application No.**
sz9-97-003	PCT/IB97/00394

The applicant is hereby notified that the International Bureau has received the record copy of the international application as detailed below.

Name(s) of the applicant(s) and State(s) for which they are applicants:

INTERNATIONAL BUSINESS MACHINES (for all designated States except US)
LUIJTEN, Ronald et al (for US)

International filing date	:	10 April 1997 (10.04.97)
Priority date(s) claimed	:	
Date of receipt of the record copy by the International Bureau	:	11 April 1997 (11.04.97)
List of designated Offices	:	

EP :AT,BE,CH,DE,DK,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE
National :BR,CA,CN,JP,KR,US

ATTENTION

The applicant should carefully check the data appearing in this Notification. In case of any discrepancy between these data and the indications in the international application, the applicant should immediately inform the International Bureau.

In addition, the applicant's attention is drawn to the information contained in the Annex, relating to:

- [X] time limits for entry into the national phase;
- [X] confirmation of precautionary designations;
- [ ] requirements regarding priority documents.

A copy of this Notification is being sent to the receiving Office and to the International Searching Authority.

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 740.14.35

Authorized officer:

Céline Faust

Telephone No. (41-22) 730.91.11

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference	**FOR FURTHER ACTION** see Notification of Transmittal of International Search Report (Form PCT/ISA/220) as well as, where applicable, item 5 below.	
SZ9-97-003		
International application No.	International filing date *(day/month/year)*	(Earliest) Priority Date *(day/month/year)*
PCT/IB 97/00394	10/04/1997	
Applicant		
INTERNATIONAL BUSINESS MACHINES CORPORATION et al.		

This International Search Report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This International Search Report consists of a total of 3 sheets.

[X] It is also accompanied by a copy of each prior art document cited in this report.

1. [ ] **Certain claims were found unsearchable** (see Box I).

2. [ ] **Unity of invention is lacking** (see Box II).

3. [ ] The international application contains disclosure of a **nucleotide and/or amino acid sequence listing** and the international search was carried out on the basis of the sequence listing
   - [ ] filed with the international application.
   - [ ] furnished by the applicant separately from the international application,
     - [ ] but not accompanied by a statement to the effect that it did not include matter going beyond the disclosure in the international application as filed.
   - [ ] Transcribed by this Authority

4. With regard to the **title,**
   - [ ] the text is approved as submitted by the applicant
   - [X] the text has been established by this Authority to read as follows:

   METHOD AND MEANS FOR DETERMINING THE USED BANDWIDTH ON A CONNECTION

5. With regard to the **abstract,**
   - [X] the text is approved as submitted by the applicant
   - [ ] the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box III. The applicant may, within one month from the date of mailing of this International Search Report, submit comments to this Authority.

6. The figure of the **drawings** to be published with the abstract is:

   Figure No. 1
   - [X] as suggested by the applicant.
   - [ ] because the applicant failed to suggest a figure.
   - [ ] because this figure better characterizes the invention.
   - [ ] None of the figures.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/IB 97/00394

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04Q11/04 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04Q H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SHIMOKOSHI K: "EVALUATION OF POLICING MECHANISMS FOR ATM NETWORKS" IEICE TRANSACTIONS ON COMMUNICATIONS, vol. E76-B, no. 11, 1 November 1993, pages 1341-1351, XP000425067 see page 1345, left-hand column, line 10 - page 1346, right-hand column, line 34 --- -/--	1,12

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
2 December 1997	17/12/1997

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Veen, G

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	KIYOSHI SHIMOKOSHI: "PERFORMANCE COMPARISON OF BANDWIDTH ALLOCATION MECHANISMS FOR LAN/MAN INTERWORKING THROUGH AN ATM NETWORK" SERVING HUMANITY THROUGH COMMUNICATIONS. SUPERCOM/ICC, NEW ORLEANS, MAY 1 - 5, 1994, vol. VOL. 3, no. -, 1 May 1994, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, pages 1405-1411, XP000438728 see page 1406, left-hand column, line 40 - right-hand column, line 30 ---	1,12
A	GONG F ET AL: "STUDY OF A TWO-LEVEL FLOW CONTROL SCHEME AND BUFFERING STRATEGIES" PROCEEDINGS OF THE CONFERENCE ON COMPUTER COMMUNICATIONS (INFOCOM), TORONTO, JUNE 12 - 16, 1994, vol. 3, 12 June 1994, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, pages 1224-1233, XP000496585 see page 1226, right-hand column, line 10 - line 55 -----	1,12

1

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
H04Q 11/04, H04L 12/56

A1

(11) International Publication Number: WO 98/46041

(43) International Publication Date: 15 October 1998 (15.10.98)

(21) International Application Number: PCT/IB97/00394

(22) International Filing Date: 10 April 1997 (10.04.97)

(71) Applicant *(for all designated States except US)*: INTERNATIONAL BUSINESS MACHINES CORPORATION [US/US]; Old Orchard Road, Armonk, NY 10504 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: LUIJTEN, Ronald [NL/CH]; Alte Landstrasse 150, CH–8800 Thalwil (CH). VU, Ken [US/US]; 219 Trillingham Lane, Cary, NC 27513 (US).

(74) Agent: KLETT, Peter, Michael; International Business Machines Corporation, Saeumerstrasse 4, CH–8803 Rueschlikon (CH).

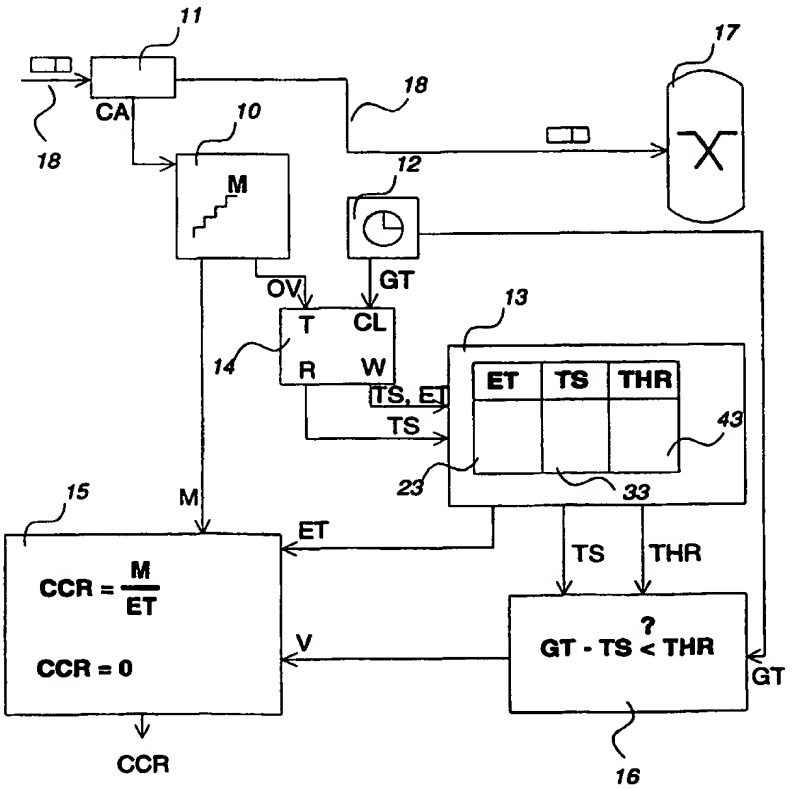
(81) Designated States: JP, KR, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND MEANS FOR DETERMINING THE USED BANDWIDTH ON A CONNECTION

(57) Abstract

The invention relates to a method for determining the use bandwidth (CCR) on a connection (18) on which information–carrying units are transported. The duration (ET) of an arrival period, during which a predetermined number (M) of the cells arrives at a certain point of the connection (18), is measured and stored. The bandwidth (CCR) at an arbitrary point of time (GT) is set to the predetermined number (M) per the stored duration (ET), if the point of time (TS) when the duration (ET) was stored is not longer ago than a predetermined threshold time interval (THR) at the arbitrary point of time (GT).



This Page Blank (uspto)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD AND MEANS FOR DETERMINING THE USED BANDWIDTH ON A CONNECTION

TECHNICAL FIELD

The invention relates to a method for determining the used bandwidth on a connection and a bandwidth determination means, particularly for the use in cell-based information-transmitting systems, such as ATM systems or switches.

BACKGROUND OF THE INVENTION

In packet-switching systems, e.g. ATM systems, several users are connected via one or more ATM switches. The users exchange information which is transmitted in form of fixed-size packets, also called cells. Since the ATM switch is limited in its capacity to handle cells without loss of information and required to avoid loss of information for certain connections, the cell rate or bandwidth which is allowed to reach the switch is limited. In some ATM systems all users are assigned an equal share of the totally available bandwidth as maximum bandwidth; in other systems users can get different shares assigned. In ATM switches with ABR (Available Bit Rate) service, the unused bandwidth is distributed over the ABR-connections.

Determining the actually used bandwidth is an important aspect for such systems since a deviation from an assigned maximum bandwidth can have various effects: Users may be confronted with additional fees for their assigned bandwidth limit being exceeded. Furthermore, free, unused bandwidth may be at least temporarily assigned to a user who exceeds his assigned bandwidth limit. Finally, queues for arriving cells need to be adapted to the used bandwidth. For all these actions, a precise determination of used bandwidth is of importance.

Two methods are known for determining the actually used bandwidth on a connection.

A first method uses an arrangement that counts the number of arriving cells for a fixed amount of time. The problem is that usually with this method one counter and one timer for each connection is to be established. Furthermore, particularly in the realistic case of a big number of connections, many results of many counting processes have to be stored in very

short time, even when less timers are provided. In worst case this may even be simultaneously. Such storage mechanism requires an appropriate provision of reliable and fast resources. This leads to expensive arrangements which still may have only limited capacity. Also, a huge amount of data is to be stored even if no cell traffic is pertinent. Another point is that the timers need to be accessed by two asynchronous processes, namely the cell-receiving process to increment the counter and a timer-based process to set the counter to zero and to read out its value. Such a twofold asynchronously controllable timer needs to be realized with a dual-port RAM which is more expensive than a single port RAM for only one asynchronous access. Also, a queuing mechanism must be implemented between the timer expiration process and the dual-port RAM containing the counters for the case when several timers expire at the same time. This again can only be avoided by providing additional, expensive hardware.

A second method counts a fixed number of cells and measures the elapsed time for these cells. It is here more probable that the storing processes when the number of cells is reached for several connections, do not occur simultaneously. A problem occurs when the number of sent cells remains below the fixed number or if the cell rate falls from a high value to a very low value during a counting process. Then the counting process does not get to an end and no terminating event stops the counting or it takes a very long time until the counting process reaches the fixed number of cells. If this happens, the last value that was used to determine the used bandwidth does no longer represent the correct value of the used bandwidth.

OBJECT AND ADVANTAGES OF THE INVENTION

The following statements refer to fixed-length cells as example. However the invention is also generally applicable for any countable information-carrying unit, such as bits or bytes belonging to a variable-length packet.

It is an object of the invention to provide a method and means for determining the used bandwidth on a connection in a more precise manner. It is another object to provide such a method and means which requires less hardware and less software for achieving a very precise result of the determination. Generally, it is an object to provide a method and means for

determining the used bandwidth on a connection which avoids the above described problems of the state of the art.

The method and means for determining the used bandwidth on a connection according to claim 1 shows the advantage that the actual value of the used bandwidth is no longer overestimated which results in more accurate bandwidth determination and as a consequence in a fairer treatment of cell traffic on different connections. The determined bandwidth is more accurate an any arbitrary point of time.

Another advantage is that the necessity of providing a separate timer per connection is avoided.

Furthermore, the invention is advantageous in that idle connections, i.e. connections that do not have cell traffic, do not initiate unnecessary table updates.

Using a predetermined threshold interval for determining how long ago a stored duration of an arrival period has been stored, is a simple and hence easy-to-implement solution for the decision whether to use the stored duration value for determining the bandwidth or not.

In the dependent claims various modifications and improvements of the method and means for determining the used bandwidth on a connection are contained.

Setting the bandwidth value to a fixed value like zero if the point of time when the stored duration was stored is at the arbitrary point of time older than the predetermined threshold time interval brings the advantage that no complicated algorithm need be used for calculating the bandwidth value and that the value zero is a good approach to any possible real value. Generally, also other fixed values than zero can be chosen. It is even thinkable that not a fixed value but a different formula for determining the used bandwidth is used. The choice mainly depends on the choice of the duration of the threshold time interval.

When the point of time when the number of the arriving cells reach the predetermined number, is stored, this time can not only be used for a calculation of the then valid bandwidth value but also for any following calculation, particularly for the next bandwidth determination process.

Using the point of time when the number of the arriving cells reached the predetermined number together with the arbitrary point of time and the predetermined threshold time

interval to determine whether the stored duration is at the arbitrary point of time older or not older than predetermined threshold time interval, is advantageous because with a minimum of stored values already a decision can be made if the stored values can be used for the bandwidth determination process or not.

When the value of the predetermined threshold time interval is stored, this value can be used for several bandwidth determination processes and is still variable such that it can be changed for achieving a different result of the bandwidth determination processes. Such change may particularly be done when the cell traffic density is changing over a larger period of time and for initializing and/or optimizing the determination process.

Since bandwidth changes with time it is useful and particularly of advantage when after storing the duration of the arrival period, the duration of the next arrival period is measured and stored and the bandwidth at the arbitrary point of time is set to the predetermined number of the cells per the last-stored duration, if the point of time when the last-stored duration was stored is at the arbitrary point of time not older than the predetermined threshold time interval . Then, for determining the actual value for the used bandwidth, the stored values which serve as basis for the determination process, are updated each time when a more actual set of values is created.

A minimum of storage space is needed if the duration of the previous arrival period is erased or overwritten when the duration of the next arrival period is stored.

No timer is needed for every connection if the point of time when the number of the arriving cells reached the predetermined number, is used together with the point of time when the number of the arriving cells reached the predetermined number the last time, to determine the duration of the arrival period.

The advantage that less storage space is needed can here be achieved by erasing or overwriting the previous point of time when the number of the arriving cells reached the predetermined number by the point of time when the number of the arriving cells reach the predetermined number again.

A very easy-to-implement and low-price storage section can be used if one stores for the connection the predetermined threshold time interval, the last-stored duration and the point

of time when the number of the arriving cells reached the predetermined number the last time in a storing means.

The invention shows its advantages particularly if it is used for a plurality of connections, since then the little equipment that is needed for its implementation makes the method and means according to the invention much cheaper than known implementations.

SUMMARY OF THE INVENTION

The invented method for determining the used bandwidth on a connection and the bandwidth determination means are applicable for a method where a fixed number of arriving information-carrying units is counted. The arrival period of these units is measured and stored. At an arbitrary point of time the used bandwidth is calculated as being the fixed number of units divided by the value of the stored duration. This calculation is only defined as the actual value of the used bandwidth if between the point of time when the duration was stored and the arbitrary point of time when the bandwidth is to be calculated, a time-difference is not exceeded. The time difference is defined as a predetermined threshold time interval.

The problem from the state of the art that a previously determined value of the used bandwidth is still considered valid, although the cell traffic on the connection may even have ended is hence solved in that the value of the bandwidth for a connection is tested for its actuality.

DESCRIPTION OF THE DRAWINGS

Examples of the invention are depicted in the drawings and described in detail below by way of example. It is shown in

Fig. 1: a block-diagram of a bandwidth determination means according to the invention,

Fig. 2: an example of a time schedule for cells arriving at a connection.

All the figures are for sake of clarity not shown in real dimensions, nor are the relations between the dimensions shown in a realistic scale.

DETAILED DESCRIPTION OF THE INVENTION

In the following, the various exemplary embodiments of the invention are described.

In figure 1 a schematic picture of a bandwidth determination means is depicted. It comprises a cell arrival signalisation means 11 which is connected to the input of a counting means 10 and which is situated inmidst a connection 18 that leads to a switching means 17. A time-calculating means 14 is connected via its first input port T to the counting means 10 and via its second input port CL to a timing means 12 which is also connected to an input of a decision means 16. Between the time-calculating means 14 and the decision means 16 is arranged a storing means 13 which has an output line leading to an input of a definition means 15. One line is leading from the storing means 13 to a read input R of the time-calculating means 14 and one line is leading from a write output W of the time-calculating means 14 to the storing means 13. Another input of the definition means 15 is connected to the counting means 10 and a third input of the definition means 15 is connected to the decision means 16. The definition means 15 has one output line.

Cells that contain information are arriving on the connection 18 at the cell arrival signalisation means 11 and are led through it towards the switching means 17. The cell arrival signalisation means 11 creates a cell arrival pulse signal CA, each time one cell arrives. The cell arrival pulse signal CA is used by the counting means 10 to increment a contained counter which counts until it reaches a predetermined number M. When the number M is reached, the counter is automatically reset to zero and begins counting again. Further, the counter then issues an overflow signal OV which enters the time-calculating means 14 at its first input port T, also called trigger input T, and thereby initiates a process in the time-calculating means 14 to determine the point of time when the overflow of the counter occurred. Therefor, the time-calculating means 14 uses the value of a global time signal GT delivered by the timing means 12, e.g. a system clock, which value is read into the time-calculating means 14 when the trigger input T receives the overflow signal OV. The result is an overflow moment

GTO in the time-calculating means 14. The overflow moment GTO when the counter reached the predetermined number M is hence determined in the time-calculating means 14.

In the storing means 13 a table is contained which has three columns: a first column 23 which is dedicated for an elapsed time ET, a second column 33 which is dedicated for a timestamp TS and a third column 43 which is dedicated for a threshold time interval THR. Since in this example only one connection 18 is discussed, the table contains only one row.

The time-calculating means 14 receives via its read input R the value that is contained in the second column 33, namely the stored value of the timestamp TS. The timestamp TS is defined as the point of time when the table was updated the last time. By subtracting this timestamp value from the value of the overflow moment GTO, the time-calculating means 14 calculates the value of the elapsed time ET which it afterwards sends via its write output W to the storing means 13 where the calculated value of the elapsed time ET is stored in the table. The elapsed time ET is hence the duration of the arrival period for a number M of cells arriving at a certain point of the connection 18, namely at the cell arrival signalisation means 11.

Then, also the value of the timestamp TS in the table is replaced by the value of the overflow moment GTO from the time-calculating means 14 in that this value is sent via the write output W to the storing means 13.

For determining at an arbitrary point in time the used bandwidth CCR for the connection 18, the decision means 16 reads the values of the timestamp TS and of the threshold time interval THR from the table and further uses the actual time which it gets as the global time signal GT from the timing means 12. By subtracting the read value of the timestamp TS from the global time signal GT which hence represents the arbitrary point of time and by comparing the result of the subtraction with the value of the threshold time interval THR the decision means 16 determines if the timestamp TS is or is not longer ago from the arbitrary point of time GT than the threshold time interval THR. With other words it is tested whether GT-TS < THR or GT-TS > THR.. The case when both values are equal can be assigned to one of the equations, whatever is preferred.

Since the timestamp TS is the point of time when the table was last updated, this test means that the decision means 16 finds out whether the stored value of the duration ET is at the

arbitrary point of time GT older than the predetermined threshold period THR or not. The result V of this operation is transmitted from the decision means 16 to the definition means 15.

In the definition means 15 the used bandwidth CCR is finally defined. The definition depends on the result V in the following way: In the case, the result V shows that the stored duration ET is at the arbitrary point of time GT older than the predetermined threshold period THR, the bandwidth CCR is set to zero. In the opposite case the bandwidth CCR is defined to be the predetermined number M per the stored duration or elapsed time ET. The number M the definition means 15 gets directly from the counting means 10. The definition means finally delivers the value for the used bandwidth CCR to its output.

In reality, the definition means 15 may at any time be requested to output the actually used bandwidth CCR for the connection 18. Then the definition means 15 initiates the decision means 16 to deliver an actual value of the result V.

The value of the threshold time interval THR is usually pregiven and chosen such that it leads to acceptable results of the used bandwidth determination procedure. A basis for the choice of the threshold time interval THR can be the maximum bandwidth CCR or peak cell rate PCR allocated to the connection 18.

The background of the invention is the fact that a running counting process cannot be used for determining the actual bandwidth CCR but only the last stored values deriving from the last counting process. It may happen that the used bandwidth CCR during such a counting process varies so strongly that the predetermined number M of cells is reached only after a very long period of time after the point of time when the counter started its counting process or it is even not reached at all. Then the counting process is pertaining on and on and the last stored values in the table till remain valid for a state-of-the-art bandwidth determination process which simply would use the last calculated bandwidth CCR which here is calculated from the stored values in the table. However, the last-stored values of the table do not longer represent the correct value of the used bandwidth CCR, since the long counting process apparently proves that the cell rate has decreased. After storing values deriving from a counting process, these values certainly stay valid for a certain period of time. The more time passes after the storing, the higher is the probability that the bandwidth CCR has

changed meanwhile and that the stored values are no longer representative for the actual bandwidth CCR.

To this adds the fact that when so much time has passed since the last storing procedure, this is an evidence for a pertinent counting process and for the assumption that the cell rate has decreased significantly.

With the invention, at some point in time after the last storing it is decided to not longer use the stored values for determining the actually used bandwidth CCR. This point in time is defined by the threshold time interval THR. Depending on the choice of the length of the threshold time interval THR the actual value of the used bandwidth CCR can be assumed to a low value, particularly zero.

For starting the process, an initialization may take place which gives the first values for the table. The threshold time interval THR is chosen according to a desired behavior of the bandwidth-determining means and the respective value of the threshold time interval THR is then stored in the table. This value is usually not amended during usage of the system, at least not as often as other parameters. However the value of the threshold time interval THR can be adapted to the connection 18, respectively to the expected or even experienced cell traffic on this connection 18. As first value of the timestamp TS the actual value of the global time signal GT is used and for the elapsed time ET the value of the threshold time interval THR or another high value may be stored.

While the counting process is running automatically, triggered by the arrivals of the cells, and hence the updating of the table content is triggered by the counter overflow, the determination process of the used bandwidth CCR is independent from time and from any of the previous processes. An external request for an output of the actual used bandwidth CCR on the connection 18 can come at any arbitrary time to the definition means 15 which with the described arrangement now gives at that arbitrary point of time a more accurate result.

In figure 2 an example for a time schedule on the connection 18 is given. The table is assumed to be preloaded e.g. with the initialization values. The time at which the table was last updated is the time t_1 whose value is also stored in the second column 33 as timestamp value TS_1. The counter was reset at that time t_1 and counts the predetermined number M of arriving cells. At the time t_2 the counter reaches the number M and has an overflow. The time-

This Page Blank (us

calculating means 14 is triggered to look at the global time signal GT and to store its value at that moment which gives the time t_2 as the overflow moment GTO.

Reading the timestamp value TS_1 from the table, the time-calculating means 14 calculates the duration ET of the elapsed time as t_2-t_1. This duration ET is then stored in the table and also the second column 33 is updated by overwriting the value of the timestamp TS_1 with a new value of the timestamp $TS_2 = t_2$.

At an arbitrary point of time GT the definition means 15 is for example asked about the actually used bandwidth CCR on the connection 18. Therefor the decision means 16 reads the value of the timestamp TS_2 and the value of the threshold time interval THR from the table. The subsequent calculation checks whether $GT\text{-}TS_2 = t_3 < THR$. Since in the depicted case $t_3 > t_2$, the stored value of the duration ET of the elapsed time is considered as too old and not longer applicable. The result V reaching the definition means 15 triggers the definition means 15 to output the value zero for the actually used bandwidth CCR. The arbitrary points of time GT may e.g. be determined by a background process that determines the bandwidth CCR on the connection at regular time intervals.

The above facts and processes are also valid for a plurality of connections 18. The arrangement of figure 1 needs only to be slightly modified, respectively enlarged. For every connection 18 a separate cell arrival signalization means 11, counting means 10 and time-calculating means 14 should be provided. The table is only amended in that for every connection 18 a separate row is provided. Hence, for every connection 18 only three values, namely the duration ET of the elapsed time, the timestamp TS and the threshold time interval THR are contained. It may even be considered to have only one value of the threshold time interval THR for a plurality of connections 18 which means that for this group of connections 18 only one value of the threshold time interval THR needs to be stored. The threshold time interval THR may even be only one value for all connections 18 and then may not even be needed as a stored value. Using a stored value for the threshold time interval THR has the advantage that this value can be changed at any time by writing a new value to the table. This renders the process very flexible. The arbitrary points of time GT may here e.g. be determined by a background process that determines the bandwidth CCR on the connections 18 sequentially and which starts again form the beginning when reaching the last connection 18, respectively entry in the table. Another modification of the described concept

This Page Blank (usp

is the provision of several threshold time intervals THR for one connection 18. Then a stepwise and hence more differentiated assignment of the assumed bandwidth CCR is possible. Depending on in which threshold time interval THR the arbitrary point of time GT lies, a different bandwidth value can be defined in the definition means 15. These values should then best lie between the last calculated value based on the table content and zero.

It is also possible to choose not a fixed value but a different formula for determining the used bandwidth CCR. An example may be the choice of an additional factor α which is introduced in the formula CCR = M/ET which is used when the stored values are considered actual and makes it to CCR = Mα/ET. Since the actually used bandwidth CCR is lower than the last determined bandwidth CCR when the threshold time interval THR is exceeded, the additional factor α will be smaller than 1.

The depicted arrangement is only an example. The several contained functions may be combined or separated, as long as the desired result is achieved. For instance, the table can be split up and need not be provided as a whole. The third column 43 may e.g. also be assigned to the decision means 16. Definition means 15 and decision means 16 may be unified as one single means. The counting means 10 may be replaced by any equivalent means that gives a signal everytime when the predetermined number M of cells is reached.

It may also be desirable to sum up the total used bandwidth over all connections 18, using the single values of the bandwidths CCR for the connections 18. This is also possible for one or several selected subsets of the connections 18, such as a subset for only the non-ABR connections.

CLAIMS

1. Method for determining the used bandwidth (CCR) on a connection (18) on which countable information-carrying units , e.g. cells, are transported, characterized in that the duration (ET) of an arrival period, during which a predetermined number (M) of said units arrives at a certain point of said connection (18), is measured and stored and that said bandwidth (CCR) at an arbitrary point of time (GT) is set to said predetermined number (M) per said stored duration (ET), if the point of time (TS) when said duration (ET) was stored, is not longer ago than a predetermined threshold time interval (THR) at said arbitrary point of time (GT).

2. Method according to claim 1, characterized in that the bandwidth (CCR) at the arbitrary point of time (GT) is set to zero, if the point of time (TS) when the stored duration (ET) was stored is at said arbitrary point of time (GT) older than the predetermined threshold time interval (THR).

3. Method according to claim 1 or 2, characterized in that the point of time (TS) when the number of the arriving units reached said predetermined number (M) is stored.

4. Method according to one of claims 1 to 3, characterized in that the point of time (TS) when the number of the arriving units reached said predetermined number (M) is used together with the arbitrary point of time (GT) and the predetermined threshold time interval (THR) to determine whether the stored duration (ET) is at said arbitrary point of time (GT) older or not older than said predetermined threshold time interval (THR).

5. Method according to one of claims 1 to 4, characterized in that the value of the predetermined threshold time interval (THR) is stored.

6. Method according to one of claims 1 to 5, characterized in that after storing the duration (ET) of the arrival period, the duration (ET) of the next arrival period, is measured and stored and that the bandwidth (CCR) at the arbitrary point of time (GT) is set to the predetermined number (M) of the units per the last-stored duration (ET), if the point of time (TS) when the last-stored duration (ET) was stored is at said arbitrary point of time (GT) not older than the predetermined threshold time interval (THR).

7. Method according to claim 6, characterized in that when the duration (ET) of the next arrival period is stored, the duration (ET) of the previous arrival period is erased or overwritten.

8. Method according to claim 6 or 7, characterized in that the point of time (TS) when the number of the arriving units reached the predetermined number (M), is used together with the point of time (TS) when the number of said arriving units reached said predetermined number (M) the last time, to determine the duration (ET) of the arrival period.

9. Method according one of claims 6 to 8, characterized in that when the point of time (TS) when the number of the arriving units reached the predetermined number (M), is stored, the previous point of time (TS) when the number of said arriving units reached said predetermined number (M), is erased or overwritten.

10. Method according to one of claims 1 to 4, characterized in that the predetermined threshold time interval (THR), the last-stored duration (ET) and the point of time (TS) when the number of the arriving units reached the predetermined number (M) the last time is stored for the connection (18) in a storing means (13).

11. Method according to claim 1 characterized in that it is used for a plurality of connections (18).

12. Bandwidth determination means comprising measuring means (10, 14) for measuring the duration (ET) of an arrival period, during which a predetermined number (M) of information-carrying units arrives at a certain point of a connection (18), and storing means (13) for storing said duration (ET) and definition means (15) for setting said bandwidth (CCR) at an arbitrary point of time (GT) to said predetermined number (M) of said units per said stored duration (ET), if the point of time (TS) when said stored duration (ET) was stored, is not older than a predetermined threshold time interval (THR) at said arbitrary point of time (GT).

13. Bandwidth determination means according to claim 12, further comprising decision means (16) for determining whether the stored duration (ET) is at the arbitrary point of time (GT) older or not older than the predetermined threshold time interval (THR) by using the point of time (TS) when the number of the arriving units reached the predetermined number (M) together with the arbitrary point of time (GT) and said predetermined threshold time interval (THR).

14. Bandwidth determination means according to claim 12 or 13, characterized in that the measuring means (10, 14) is designed to measure, after storing the duration (ET) of the arrival period, the duration (ET) of the next arrival period, and that the storing means (13) is designed to store said duration (ET) of said next arrival period and that the definition means (15) is designed such that the bandwidth (CCR) at the arbitrary point of time (GT) is set to the predetermined number (M) of the units per the last-stored duration (ET), if the point of time (TS) when the last-stored duration (ET) was stored is at said arbitrary point of time (GT) not older than the predetermined threshold time interval (THR).

15. Bandwidth determination means according to one of claims 12 to 14, characterized in that the storing means (13) is designed such that when the duration (ET) of the next arrival period is stored, the duration (ET) of the previous arrival period is erased or overwritten.

16. Bandwidth determination means according to one of claims 12 to 15, characterized in that the measuring means (10, 14) is designed to use the point of time (TS) when the number of the arriving units reached the predetermined number (M), together with the point of time (TS) when the number of said arriving units reached said predetermined number (M) the last time, to determine the duration (ET) of the arrival period.

17. Bandwidth determination means according to one of claims 12 to 16, characterized in that the storing means (13) is designed such that when the point of time (TS) when the number of the arriving units reached the predetermined number (M), is stored, the previous point of time (TS) when the number of said arriving units reached said predetermined number (M), is erased or overwritten.

18. Bandwidth determination means according to one of claims 12 to 17, characterized in that the storing means (13) is designed such that the predetermined threshold time interval (THR), the last-stored duration (ET) and the point of time (TS) when the number of the arriving units reached the predetermined number (M) the last time is stored for the connection (18).

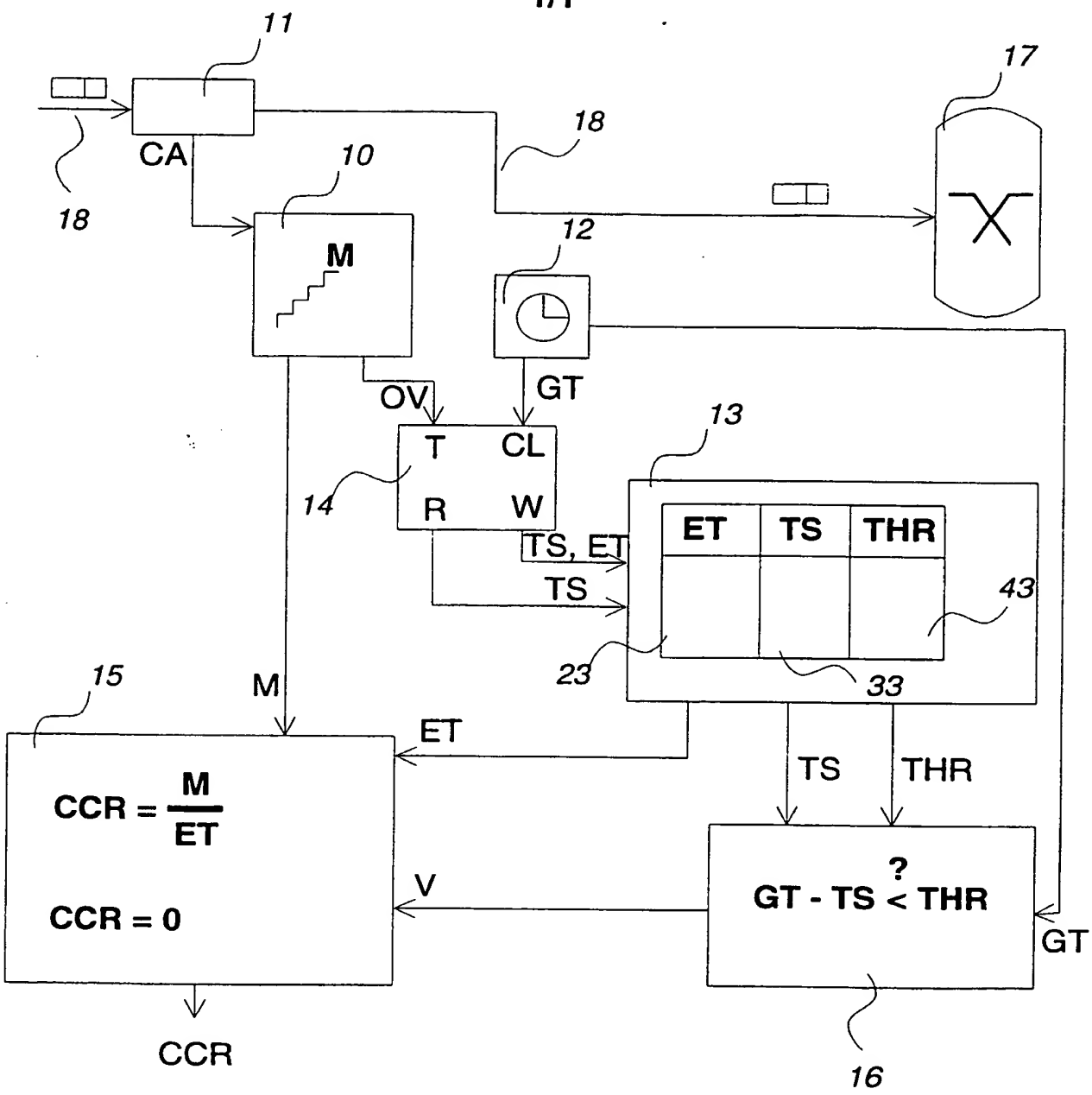

11
17
18
CA
10
18
M
12
GT
OV
T
CL
R
W
14
13
ET
TS
THR
TS, ET
TS
43
23
33
15
M
ET
TS
THR
CCR = M/ET
CCR = 0
V
GT - TS < THR ?
GT
CCR
16

Fig. 1

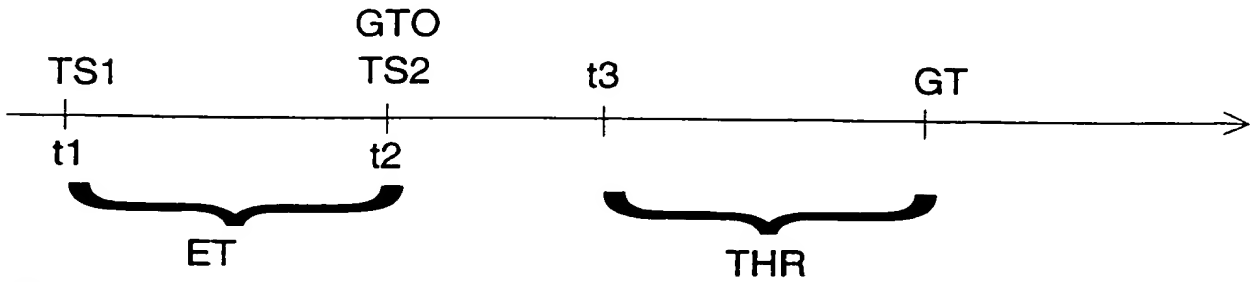

GTO
TS1
TS2
t3
GT
t1
t2
ET
THR

Fig. 2

PUBLICATION DATE: 01. 05. 94 (further bibliographic data on next page)

Performance Comparison of Bandwidth Allocation Mechanisms for LAN / MAN Interworking through an ATM Network

Kiyoshi Shimokoshi

Institute of Communications Switching and Data Technics
University of Stuttgart, F.R.Germany

ABSTRACT : Recently LAN/MAN interconnection through ATM networks has attracted attention as an application in the first stage of B-ISDN era. In this paper the interworking methods for CL data transfer through an ATM based B-ISDN, which can be applied both for i) indirect, implicit service and ii) direct, explicit service, are evaluated. The paper offers a quantitative assessment on some major algorithms to allocate effectively the network resources to the CL traffic. Performance comparison of six bandwidth allocation methods based on the BR (Bandwidth Renegotiation), VCE (VC Establishment) and FBA (Fixed Bandwidth Allocation) which are combined with TEF (Traffic Enforcement Function) mechanisms will be given by a simulation approach.

1. Introduction

In recent years, various LANs and MANs have been developed and spread to communicate between fully distributed computers among office users [1]. On the other hand, ATM technology is expected for an infrastructure of B-ISDN and its standardization is currently progressed rapidly by ITU-TS. The LAN/MAN interconnection through ATM networks has attracted attention as an application in the first stage of B-ISDN era and as a driving-force for advancement of the ATM networks. Namely the use of ATM networks as Wide Area Networks (WANs) is expected strongly, because LAN/MAN data transfer characteristic which has burstiness, is applicable to distinctive features of the cell based ATM. Here it should be noticed that LANs and MANs are connectionless (CL), while ATM based B-ISDN offer basically connection-oriented (CO) services. Although for the CL service supporting through the ATM network an efficient resource allocation function to the CL data traffic is necessary to avoid insufficient use of network resources, the traffic aspects for the CL services are still under study.

In this paper, performance comparison of interworking mechanisms by using a simulation technique concerning with cell buffer size and the number of CL sources will be given, to select a most suitable method for efficiently allocating the bandwidth to CL data traffic. Six bandwidth allocation methods based on Bandwidth Renegotiation (BR), Fixed Bandwidth Allocation (FBA) and Virtual Connection Establishment (VCE) are evaluated with 4 types of Traffic Enforcement Function (TEF) of policing based mechanisms [5]. In Ref. [6] we also evaluated the transient behavior of the methods and influence of the LAN/MAN peak rate, packet length, request message delay and cell loss control methods when a packet is discarded on the interworking performance.

The author is now with OKI Electric Industry Co., Ltd., 1-3-B23 Nakase, Mihama-ku, Chiba-shi, Chiba 261-01, Japan.

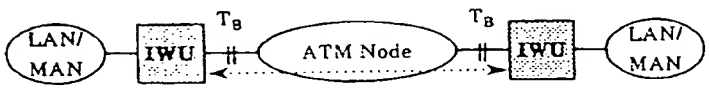
2. LAN / MAN Interworking Issues

To support LAN/MAN interconnection through an ATM network, there are two types of methods, i.e., i) end-to-end B-ISDN connection (indirect, implicit method) and ii) B-ISDN CLSFs (direct, explicit method) [2]-[4][7]-[10], which are depicted in Fig.1. The IWUs convert the protocols between LANs/MANs and ATM networks and translate LAN/MAN local addresses from/to E.164 network addresses. In case of type i), the IWUs are connected by VCs (Virtual Connections), PVCs (Permanent VCs) or semi-PVCs[15] and it may cause waste bandwidth if PVCs with peak rate are adopted, due to burstiness of the CL data traffic. Since this manner, the indirect method may be applied only for transient phase of the network evolution and for small closed networks like private networks where there are small number of IWUs [11]. In service type ii), CLSFs which are located within B-ISDN or outside B-ISDN terminate CL protocols and route cells to destination address according to routing information included in user data. In order to reduce the number of CLSFs and to concentrate the CL traffic, it is supposed to locate the CLSFs hierarchically [10]. The recommendations [2]-[4] illustrate that IWU-CLSF links can be connected by VCs, PVCs or semi-PVCs, while only semi-PVCs can be used between CLSFs, i.e., a kind of virtual overlay network between CLSFs is established on top of the ATM network. The direct method may provide efficient use of the network resource by means of statistical multiplexing of CL data traffic. Hence it may be useful for wide spread networks such as public networks where there is a great demand to interconnect LANs/MANs through ATM networks. For both service i) and ii), a dynamic bandwidth allocation mechanism is required to use effectively the network resource and to ensure the data transfer quality [7]-[13] for IWU-IWU and IWU-CLSF connections.

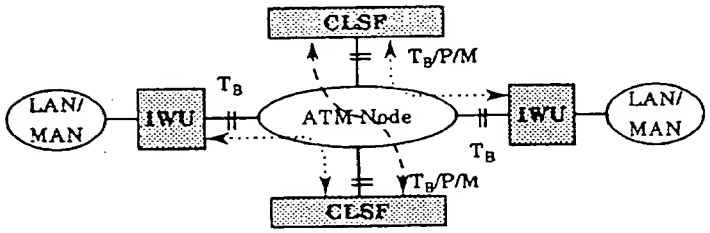
3. Interworking Mechanisms

3.1 Traffic Enforcement Function (TEF)

The output cell traffic rate from IWUs/CLSFs is regulated by the TEF mechanism to be below the assigned bit rate. In other words, a set of the TEF and IWU/CLSF buffer can realize a kind of traffic shaping function. The TEFs are implemented based on UPC (Usage Parameter Control), i.e., policing algorithm such as Leaky Bucket (LB), Jumping Window (JW), Moving Window (MW) or Pseudo Jumping Window (PJW) [5]. In Ref.[5] 5 types of UPC algorithm have been evaluated on the performance and implementation complexity. In this paper the above 4 algorithms are selected because of their easiness for hardware implementation. Figs.2 (a) - (d) illustrate functional models of the TEF mechanisms.



i) Indirect Service



ii) Direct Service with CLSFs

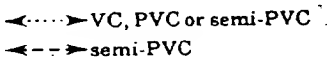


Fig.1 LAN/MAN Interconnection through an ATM Network

LB Based Mechanism

The LB is a credit (token) system which has a token pool [14]. K tokens are generated every D [cell slots] and stored in the token pool with size B. If there are less than R tokens in the pool when a cell arrives, the cell is queued in a cell buffer as long as there are queueing places. If there are R or more than R tokens in the pool, the cell is transmitted and the number of tokens is reduced by R. The stored cells can be transmitted with link speed when the LB get new tokens and the number of pooled tokens is R or more than R. These parameters indicate that assigned cell rate R_C and maximum burst size S_B are

$$R_C = \frac{K}{DR}, \qquad S_B = \frac{B+K}{R}, \tag{1}$$

where $B/R < D$ for $D > 1$ and $K < R$ for $D = 1$ in order to restrict the bit rate. Furthermore, the LB control the regulate interdeparture time of two consecutive cells when $B = R = K = 1$.

JW Based Mechanism

The JW can be defined as a counting algorithm [5] that counts arriving cells in a fixed time interval (window) T [cell slots]. N [cells] are allowed in each window. A new window starts immediately when the preceding one terminates. The counter value is reset to zero at the beginning of each window. If the counter value is equal to N, the arriving cell is queued in the cell buffer as long as there are free queueing places. N queued cells can be sent out with link speed when a new window starts. In JW the followings are obvious,

$$R_C = \frac{N}{T}, \qquad S_B = 2N. \tag{2}$$

MW Based Mechanism

The MW has same parameters T [cell slots] and N [cells] as the JW [5]. The difference from the JW is that logically separated windows start at every cell slot. An arrived cell can be sent only if all the counter values of the T windows are below the threshold N. If one or more than one counter values are equal to N, arriving cells are queued in the cell buffer or discarded immediately if there is no queuing places available. A stored cell can be transferred after a window ends. The TEF based on the MW can be characterized as follows,

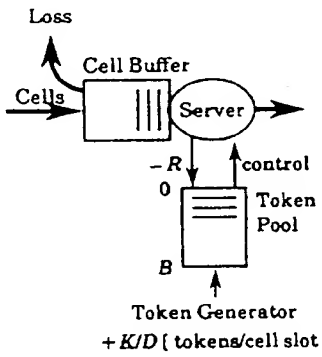
$$R_C = \frac{N}{T}, \qquad S_B = N. \tag{3}$$

PJW Based Mechanism

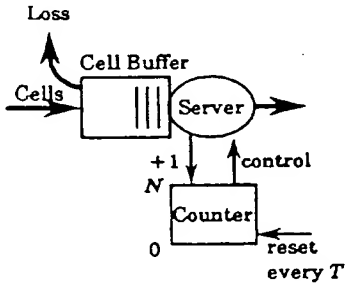
The PJW which has three parameters Na, Np and T, is an advanced mechanism of the JW of which $N = Na$ [5]. The windows of the PJW appear in the same manner as the JW. The parameter Na shows maximum number of allowed cells in a window and Np indicates the pseudo cell queueing places (credits) which means that even if the number of arrived cells exceeds Na, the cells can be sent as long as the counter value is less than $(Na + Np)$. Cells arrived at when the counter value is equal to $(Na + Np)$ are stored in the cell buffer only if there are free queueing places. Then after the window ends, the pseudo queued cells are served, i.e., the cell counter value is decrementted by Na, and queued cells can be sent out with link speed as long as the counter value is less than $(Na + Np)$. The following relations can be easily derived from the definition of the PJW,

$$R_C \sim \frac{Na}{T}, \qquad S_B = 2Na + Np. \tag{4}$$

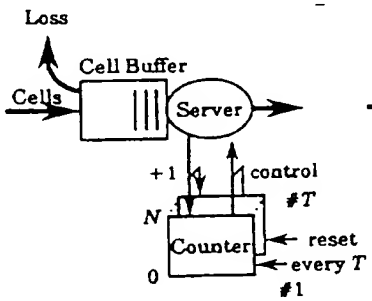
Ref. [5] shows that the PJW can be expected as one of the most suitable mean rate UPC mechanisms from view points of policing accuracy for compliant traffic which suffers long-term Cell Delay Variation (CDV), quick response ability to non-compliant cells and easiness of the implementation.



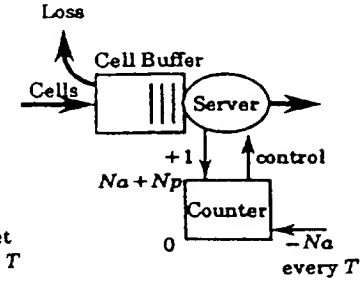
(a) LB Based Mechanism



(b) JW Based Mechanism



(c) MW Based Mechanism



(d) PJW Based Mechanism

Fig.2 Functional Models of the TEF Mechanisms

the IWU/CLSF, a new VC with peak rate bandwidth is established after receiving the acknowledgement from the network. Then after that the packet is sent completely, i.e., EOM (End of Message) or SSM cell is sent out, the established VC is released immediately. Although most efficient use of the bandwidth may be expected in the VCE-IR, it may cause processing overload of call processors in IWUs/CLSFs and the network if a number of packets arrive concurrently.

VCE-DR (VC E - Delayed Release)

The delayed release of established VCs can decrease the call processing and overcome the above problem. After a EOM or SSM cell is sent out, the VC is released after a time interval Td. Following packets may be transferred without a new VC request. The parameter Td has to be chosen carefully, because it determines the utilization of the network resources. Obviously there is a tradeoff between the effective use of the bandwidth and signalling overhead caused by connection setup and release.

4. Simulation Model

For each CL terminal connected to LAN/MAN a hierarchical traffic model covering burst, packet and cell level is assumed [8]. Here the generation of packets is according to the well-known burst silence model. In the burst phase the number of packets are geometrically distributed with mean Ex, and their interarrival time is constant, Tp. The silence phase generates no packet and is distributed negative-exponentially (mean Es). The 2-phase model may represent packet transfer between LANs/MANs [8]. In the IWUs the packets are segmented into ATM cells. (see Fig.4)

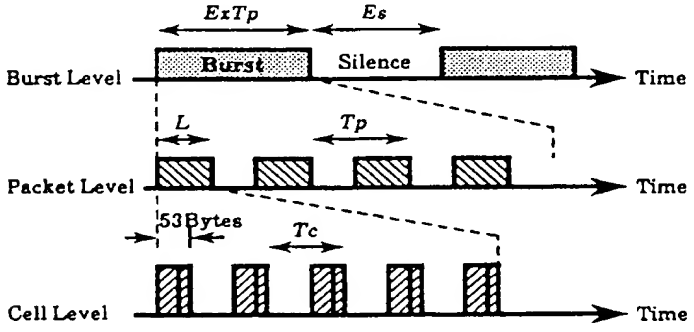


Fig.4 Hierarchical Source Traffic Model

To evaluate the interworking methods a network model which consists of a IWU and a number of CL sources is assumed as shown in Fig.5. The IWU has only cell buffer with size Q but not packet buffer and it means that arrived cells are sent immediately without the packet reassembling to ensure the high speed communication service. When an arrived cell which has ST of BOM or COM (Continuation of Message) is lost due to the cell buffer overflow, succeeding cells in the same packet are also discarded even if there are free queueing places [6]. The bandwidth increase and VC establish requests suffer a certain propagation and processing delay which is distributed negative-exponentially with mean Dm. The request messages are also rejected by probability Pr. In the IWU there is a message timer Tm which is started when a request message is sent. The IWU sends also a re-request message when it receives a negative message from the network or the timer expires.

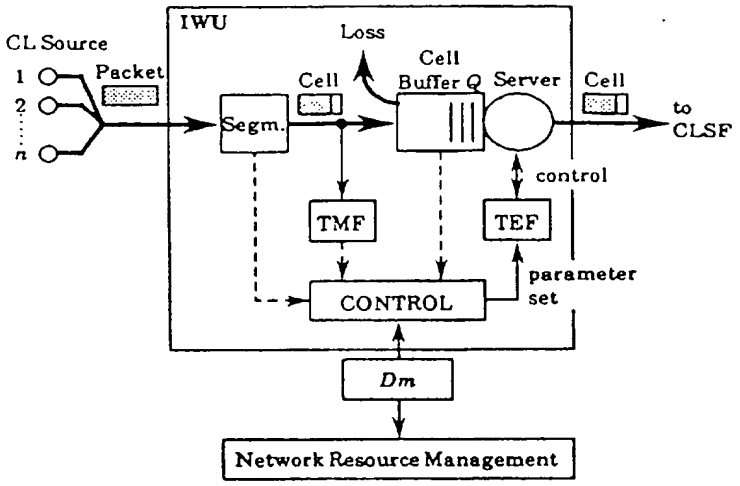


Fig.5 Simulation Network Model

To simulate the above mechanisms easily we assume the followings.

- For the dynamic bandwidth allocation of BR and VCE methods, only parameters, the token generating interval D for TEF = LB and the number of cells per window N for TEF = JW and MW and Na for TEF = PJW, can be changed. The other parameters such as K, R and B for the LB and T for the window based TEFs are fixed with initial values.
- Processing load and delay for VC establish request of the VCE methods are mostly equal to those for bandwidth increase request of the BR methods.

Furthermore, we represent the parameters of the compared interworking systems as follows.

- Interworking Methods : BR-NCB-T ($THU/Q, THL/Q$),
  BR-NCB-A ($THU/Q, THL/Q, Xi, Xd$),
  BR-TMF ($THU/@, THL/@, M$),
  FBA,
  VCE-IR, and
  VCE-DR (Td).
- TEF Mechanisms : LB (K, D, B, R),
  JW (N, T),
  MW (N, T), and
  PJW ($Na+Np, T, Np/(Na+Np)$),

where @ indicates N for TEF = JW and MW, Na for TEF = PJW, and DR/K for TEF = LB.

For a case study, the following values to the CL source parameters are assumed ; Ex = 50 [packets], Es = 3.168 [s], Tp = 7.04 [ms], L = 100 [cells] and LAN peak rate = 10 [Mb/s]. Link capacity in the network is set to 155.52 [Mb/s]. These parameters characterize each CL source as mean rate = 500 [Kb/s], minimum cell interarrival time Tc = 35.2 [μs] and mean cell interarrival time = 704 [μs]. Dm, Tm and Pr are also set to 10 [ms], 100 [ms] and 5×10^{-2}, respectively. Then we set the initial or fixed TEF parameters to LB (20, 704μs, 20, 1), JW (20, 704μs), MW (20, 704μs) and PJW (20, 281.6μs, 0.6) for the dynamical bandwidth allocation excluding the FBA. The parameters mean equivalent peak rate allocation (see equations (1) - (4)), where the window length of PJW is less than the others because the ratio $Np/(Na + Np)$ is set to 0.6 which is the most applicable value for the PJW policing [5].

3.2 Bandwidth Allocation Mechanisms

In the following as comparing interworking mechanisms, 3 Bandwidth Renegotiation (BR) methods (BR-NCB-T, BR-NCB-A and BR-TMF), 2 VC Establishment (VCE) methods (VCE-IR and VCE-DR) and Fixed Bandwidth Allocation (FBA) method are defined. Figs.3 (a) - (f) show example behaviors of the interworking mechanisms.

(1) BR Method

BR - NCB - T (BR - Number of Cells in Buffer - Two Level)

BR-NCB-T allocates either LAN/MAN peak rate or a small rate [13]. First the bandwidth is set to the minimum rate which corresponds to that for the window based mechanisms such as JW, MW and PJW only 1 cell is allowed to pass the TEF in every window. Then when the number of cells in the cell buffer exceeds a certain threshold *THU*, the bandwidth increase request is sent to the network management entity. After receiving a positive message, the assigned bandwidth is updated up to the peak rate. When the cell number is less than an another threshold *THL*, the bandwidth is set to the minimum rate again. This method needs knowledge about peak rate of the LAN/MAN.

BR - NCB - A (BR - Number of Cells in Buffer - Adaptive)

The BR-NCB-A monitors the number of stored cells in the IWU/CLSF buffer [12], like the previous scheme. When the cell number exceeds a given threshold *THU*, a bandwidth renegotiation request message is sent to the network to get extra bandwidth. Then after receiving the positive message, the bandwidth is incremented by a fixed value *Xi*. Oppositely when the number of queued cells is less than a given threshold *THL*, a fixed part of the assigned bandwidth, *Xd*, is released immediately.

BR - TMF (BR - Traffic Monitoring Function)

For the BR-TMF method, the incoming cell traffic volume is monitored directly by the TMF mechanism, which is based on UPC function in the IWUs and CLSFs [8][9]. The TMFs are also realized by LB, JW, MW and PJW based algorithms similar to the TEF, and the TEF and TMF are implemented in same algorithm. In case of TMF = JW, MW and PJW mechanisms, average number of arrived cells in last *M* windows is measured, and for TMF = LB average cell interarrival time of last *M* cell intervals is observed. When the monitored average value exceeds a certain threshold *THU*, a bandwidth enlargement request is sent. After receiving a positive message a bandwidth increase procedure is performed. Then if the monitored value is below a threshold *THL*, a part of the bandwidth is released immediately. Since the BT-TMF method measures the traffic volume directly, the bandwidth can be increased and decreased dynamically based on the measured value.

(2) VCE Method

VCE-IR (VC E - Immediate Release)

The VCE-IR scheme establishes a VC for each CL packet transfer [7][10]. More precisely, when a cell of which ST (Segment Type) is BOM (Beginning of Message) or SSM (Single Segment Message) arrives at

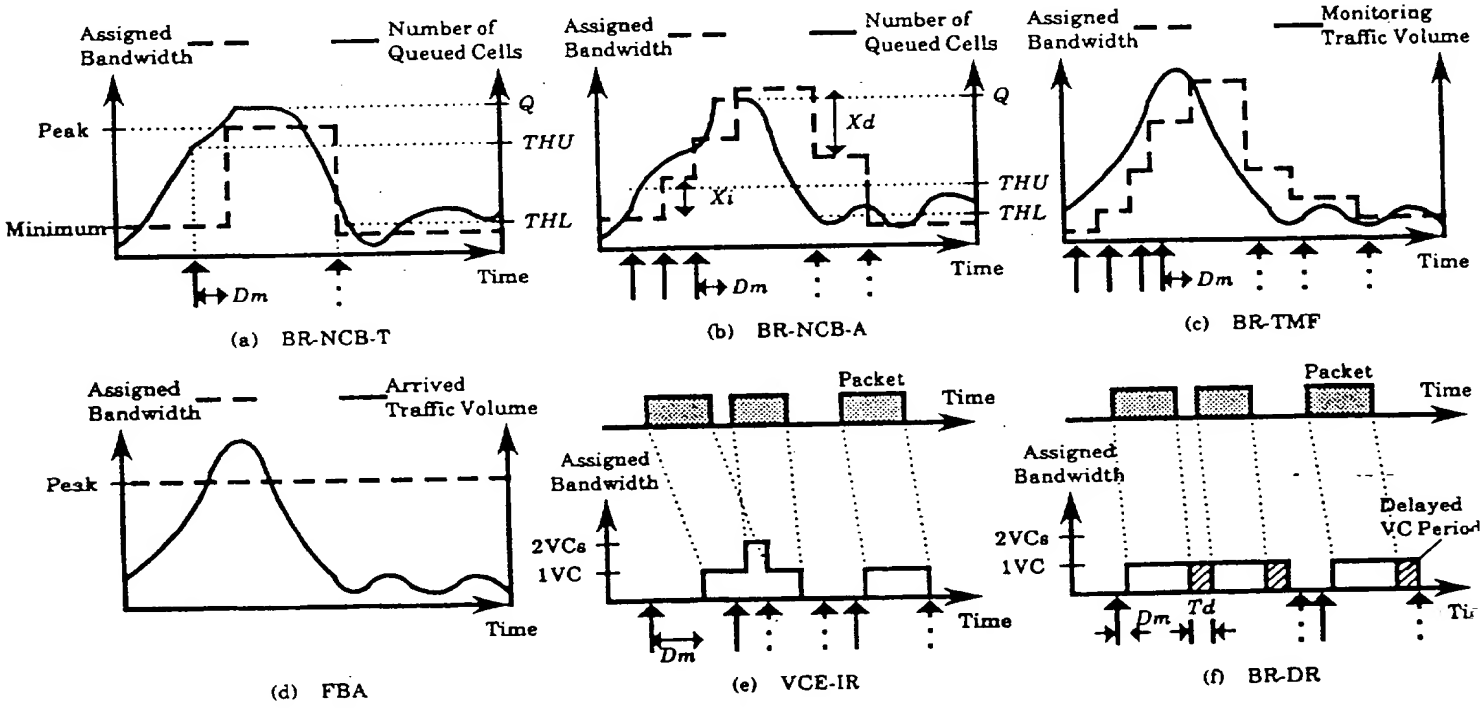


Fig.3 Example Behaviors of the Interworking Methods

[10] H.Kasahara et al., "A Study on Connection-Less Type Data Communication via an ATM Network", *IEICE Spring National Conference*, B-455, April 1990 (in Japanese).

[11] J.Cherbonnier et al., "ATM Direct Connectionless Service", *ICC93*, pp.1859-1863, Geneva, May 1993.

[12] L.Mongivoi et al., "A Proposal for Interconnecting FDDI Networks through B-ISDN", *INFOCOM91*, pp.1160-1167, Bal Harbour, FL, April 1991.

[13] G.Bianchi et al., "An Optimal Bandwidth Allocation Algorithm for Remote Bridging of FDDI Networks across B-ISDN", *GLOBECOM92*, pp.1623-1629, Orlando, FL, December 1992.

[14] H.Chao, "Design of Leaky Bucket Access Control Schemes in ATM Networks", *ICC91*, pp.180-187, Denver, CO, June 1991.

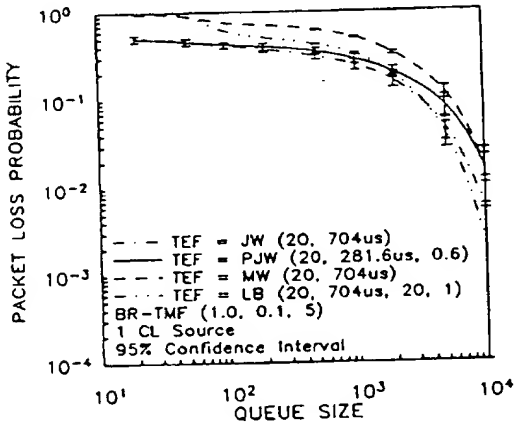


Fig.6 Comparison of the TEF Mechanisms

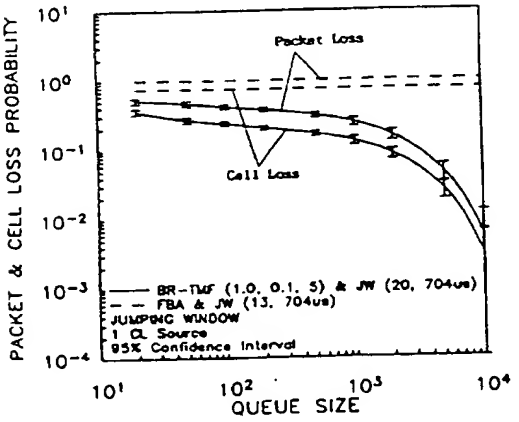
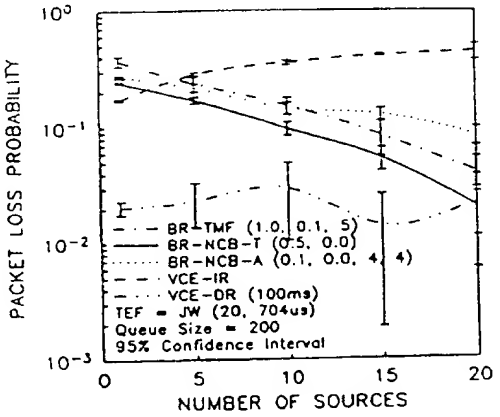
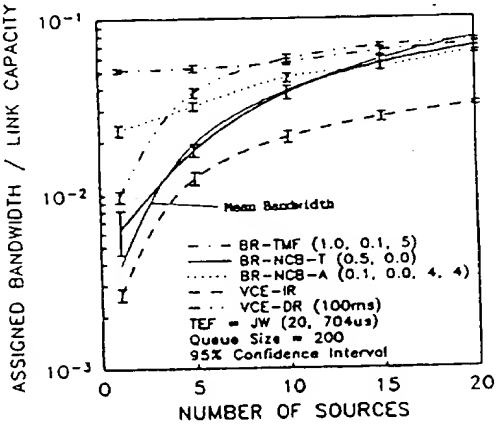


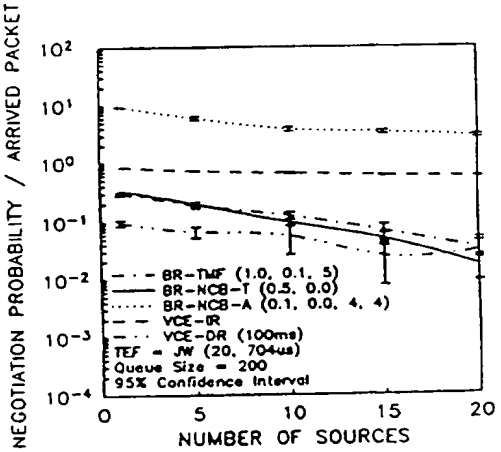
Fig.7 Performance Limitation of the FBA Method



(a) Packet Loss Probability in the IWU Buffer



(b) Assigned Bandwidth for the CL Traffic



(c) Renegotiation Probability

Fig.9 Influence of the Number of CL Sources

5. Simulation Results

Fig.6 shows performance comparison of the TEF mechanisms in case of the BR-TMF (1.0, 0.1, 5) method. The difference of the TEF mechanisms has influence only on the BR and FBA methods, because for the VCE all the TEF mechanisms allocating the peak rate are in principle equivalent. The performance largely depends on the maximum allowed burst size of the TEFs. As represented in equations (1) - (4), JW and LB can allow longest burst while the PJW follows them, then these mechanisms indicate relatively better performance with lower packet loss probability. The assigned bandwidth which does not depend on the queue size, on the other hand, is approximately 6.5 Mb/s for the JW and PJW, 5.3 Mb/s for the MW and 9.3 Mb/s for the LB. The LB based TEF monitors M cell interarrival times, so that it is rather hard to occur the bandwidth release and the result causes highest bit rate allocation.

In Fig. 7 packet and cell loss probabilities of the FBA in which the mostly same bandwidth is allocated as the average assigned one for the BR-TMF are depicted with the BR-TMF method. It is obvious that however, the FBA causes no processing load in the IWU and the network, it cannot realize the sufficient loss probability.

Figs.8 (a) - (f) are the performances of all the interworking methods except the FBA, with TEF = PJW (20, 281.6μs, 0.6) and they show the following remarks.

- From a view point of data transfer quality such as packet loss probability and cell transfer delay, the VCE-DR is most suitable method and the VCE-IR indicates also better performance in the large queue size range more than 10^2. It is because that both methods can transmit the VC establish request to get the peak rate allocation immediately, when a first cell (ST = BOM or SSM) in a packet arrives. The other BR methods, on the other hand, must wait till when the monitoring value exceeds the given threshold.
- The BR-TMF method needs large queueing capacity in the IWUs/CLSFs to achieve lower packet loss probability and it causes large transfer delay as shown in Fig.8 (b).
- From a view point of the network design, the assigned bandwidth of only the VCE-IR can be below the mean LAN/MAN bandwidth, because since the cells are stored in the cell buffer during VC establishment, the VC existing period for packet transfer is shorter than the packet length time at output of the LAN/MAN. However relatively lower bandwidth the BR-NCB-T and VCE-DR can allocate, it should be noticed that the assigned traffic volume of the latter largely depends on its delayed time interval for VC release as mentioned before.
- On the processing load of the IWUs/CLSFs and the network shown in Figs.8 (e) and (f) can be said as follows. In the small queue size range the VCE-DR indicates lowest negotiation probability for each arrived packet, while in the large queue size range the processing load of BR-NCB-T method can be decreased because of its higher threshold. The BR-TMF method shows also comparatively good performance having less relation with the queue size.

Figs. 9 (a) - (c) show the influences of the number of CL sources on interworking performance. As shown in the figure (a), only three BR methods, BR-NCB-T, BR-NCB-A and BR-TMF can reduce packet loss probability according to the increase of the number of sources. The reason is that the increase of the CL sources makes the burstiness of the traffic smaller, and it causes lower renegotiation probability to increase the bandwidth for the BR methods, but not for the VCE methods as indicated in (c). Furthermore, when the number of CL sources is equal to 20 which means LAN traffic load = 1.0 , the allocated bandwidth of the BR methods and the VCE-DR approaches also peak (i.e., mean) bandwidth. Only the VCE-IR indicates the efficient use of the resource because of its effort of queueing cells during VC establishment.

6. Conclusion

To interconnect the LANs/MANs through the ATM based B-ISDN, a problem of the bandwidth allocation for the CL traffic arise to be solved. In this paper we have evaluated some interworking methods such as the BR, FBA and VCE which are combined with the TEFs based on the LB, JW, MW and PJW mechanisms. These methods can be applied for both service type i) (indirect, implicit service) and ii) (direct, explicit service). We simulated and compared the six methods as a case study. Simulation results addressed that the VCE-DR is a most applicable to offer the required data transfer quality to users, the VCE-IR can achieve the most effective use of the network resource, and the BR-NCB-T method may cause less processing load than the other BR methods. Furthermore, the difference of the TEFs influences only on the BR and FBA, and the JW and PJW which can allow longer burst input to enter the network are relatively suitable with regard to data transfer quality and bandwidth efficiency. For actual LAN/MAN interworking, the demand to interconnect the LANs/MANs should be taken into account carefully and some kinds of combination of the methods must be considered. Selecting the most suitable combination for each network configuration is being currently studied [6].

Acknowledgement

The author would like to thank Professor Dr.-Ing.Paul J.Kühn and Mr. Uwe Briem, University of Stuttgart, F.R.Germany for their useful discussions.

References

[1] W.Lemppenau et al., "Access Protocols for High Speed LANs and MANs", *11th EFOC&N*, June 1993.

[2] *ITU-T Recommendation I.211*, "B-ISDN Service Aspects".

[3] *ITU-T Recommendation I.327*, "B-ISDN Functional Architecture".

[4] *ITU-T Draft Recommendation I.364*, "Support of Broadband Connectionless Data Service on B-ISDN", June 1992.

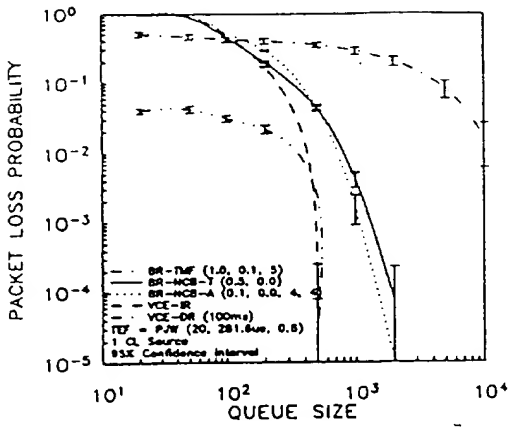
[5] K.Shimokoshi, "Evaluation of Policing Mechanisms for ATM Networks", *IEICE Transactions on Communications*, Vol.E76-B, No.11, November 1993.

[6] K.Shimokoshi, "A Simulation Study on LAN/MAN Interconnection with an ATM Network", submitted to *IEICE Transactions on Communications*.

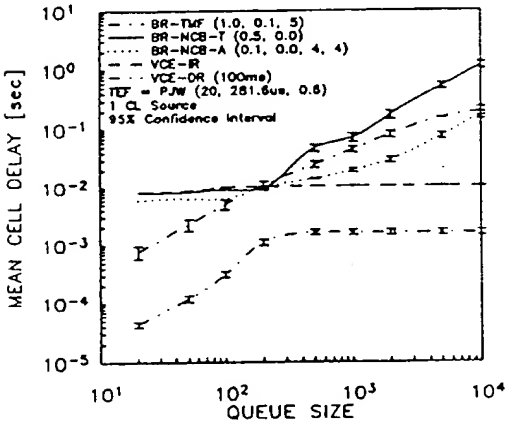
[7] P.Kühn, "Data Transfer through Interconnected Metropolitan and ATM Networks", *5th IEEE Workshop on Metropolitan Area Networks*, Taormina, Italy, May 1992.

[8] W.Schödl et al., "A Bandwidth Allocation Mechanism for LAN/MAN Interworking with an ATM Network", *5th IEEE Workshop on Metropolitan Area Networks*, Taormina, Italy, May 1992.

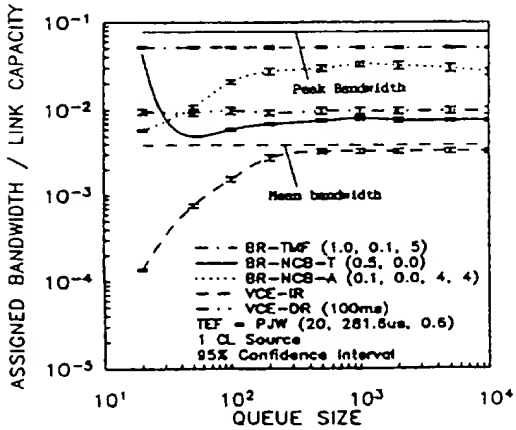
[9] M.Gerla et al. "LAN/MAN Interconnection to ATM : A Simulation Study", *INFOCOM92*, pp.2270-2279, Florence, Italy, May 1992.



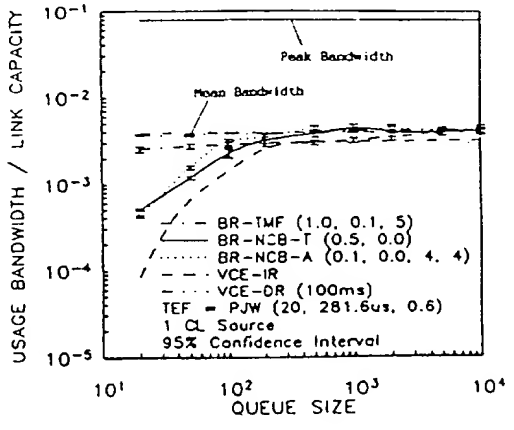
(a) Packet Loss Probabilility in the IWU Buffer



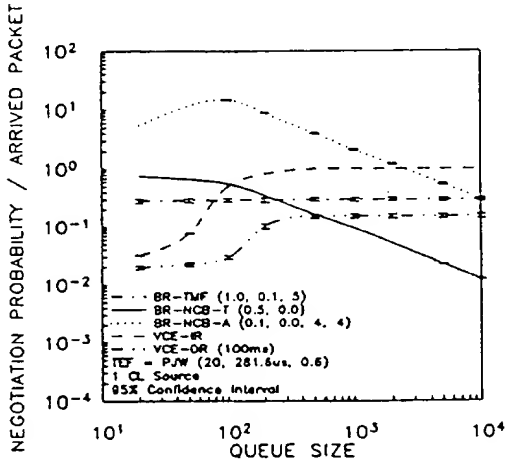
(b) Mean Cell Transfer Delay in the IWU



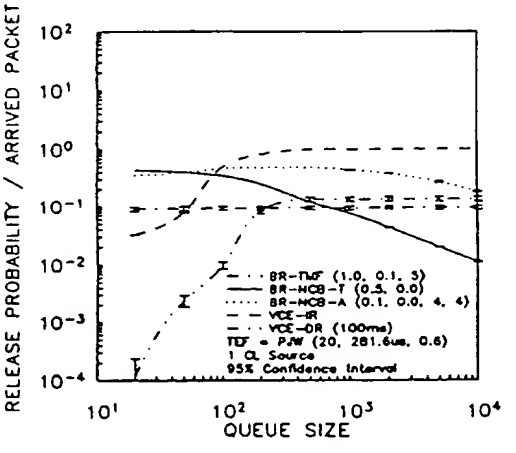
(b) Assigned Bandwidth for the CL Traffic



(d) Actual Usage Bandwidth of the CL Traffic



(e) Renegotiation Probability



(f) Bandwidth/VC Release Probability

Fig.8 Performance Comparison of the Interworking Methods

PAPER

Evaluation of Policing Mechanisms for ATM Networks

Kiyoshi SHIMOKOSHI†, *Member*

SUMMARY To realize Broadband ISDN, which provides multi-media services, ATM (Asynchronous Transfer Mode) has been standardized by CCITT and the development of the system is accelerating towards the 21st century. The packet-oriented information transfer based on fixed size blocks called cells provides a very flexible allocation of transmission capacity to different connections. On the other hand, to ensure the QoS (Quality of Service) for all established connections it is necessary to monitor and regulate the input traffic from each user based on usage parameters which are negotiated between user and network at connection set-up, i.e., a policing function is required. In this paper some requirements for a policing function will be given. Accuracy of the policing decision for violating and well-behaving sources, tolerance with respect to cell delay variation (CDV) which is caused by multiplexing functions between the source terminal and the policing device, time to detect arriving violating cells, implementation complexity, and amount, i.e., cost effectiveness, are discussed mainly. We present simulation results for five policing mechanisms, Leaky Bucket (LB), Jumping Window (JW), and Moving Window (MW) which have been already well-known, Pseudo Jumping Window (PJW), and Pseudo Moving Window (PMW) which are proposed mechanisms. PJM and PMW mechanisms required a pseudo cell buffer with finite queueing capacity to the corresponding JW and MW mechanisms, respectively. These two mechanisms can be expected as advanced methods from view points of the accuracy of the policing for long-term fluctuated compliant source, fast reaction ability and restrictness to long burst traffic comparing with the above existing methods. We compare the five mechanisms based on the above requirements and show that the PJW and the LB are the most effective mechanisms for mean rate policing in ATM networks.

key words: ATM, B-ISDN, traffic control, UPC, policing

1. Introduction

The development of the telecommunication network has been carried forward by ATMization, intellectualization, and opticalization in parallel. Further progress of these three technologies is required for the telecommunication network toward the 21st century, to provide the customer with highly developed services such as multi-media service and personal telecommunication. In those technologies, ATM (Asynchronous Transfer Mode) is expected as an infrastructure to realize the Broadband—Integrated Services Digital Network (B-ISDN), and the standardization by CCITT [1], [2] as well as the development of the node/link/terminal systems are accelerating.

ATM based on a label multiplexing principle means transfer and switching fixed length packets which are called cells. In the case of a demand service, the user sends a message to the network at the call set-up phase including the source traffic characteristics such as peak rate, average rate, burstiness, peak duration, and the required QoS (Quality of Service) class that means maximum cell loss probaliliy and maximum cell delay variation, to establish a VCC (Virtual Channel Connection). By using this information and the current load condition of the network resources, a control entity decides whether the VCC can be accepted or must be rejected. This is calld CAC (Connection Admission Control) which is one of the traffic control functions performed at the call level in ATM networks [3].

In order to protect the network resources and to ensure the negotiated QoS for all established connections, it is necessary to detect and regulate the non-compliant input traffic of malicious and misbehaved users which is sent in excess of the negotiated parameters. A so-called "Policing" or "UPC (Usage Parameter Control)" function which is a cell level traffic control function [3], has to be introduced. A policing function has to perform two types of actions; detecting a violating cell and marking (tagging) or discarding of it. Various policing mechanisms based on a Leaky Bucket principle or window algorithms have been proposed and performance comparisons have been done [4]-[11]. All mechanisms have both merits and defections.

In this paper some of the requirements for a source policing function, such as impact on non-compliant and well-behaving traffic, tolerance to cell delay variation (CDV) which is caused by multiplexing funtions located between source terminal and policing devices, time to detect violating cells, implementation complexity, and amount of it are discussed mainly. Furthermore, we propose two policing mechanisms, Pseudo Jumping Window (PJW) and Pseudo Moving Window (PMW) which have pseudo cell queueing function, and compare the performance of these mechanisms with the existing well-known Leaky Bucket

Manuscript received January 28, 1993.
Manuscript revised April 9, 1993.

† The author is with the Institute of Communications Switching and Data Technics, University of Stuttgart, Seidenstrasse 36, 70174 Stuttgart 1, F. R. Germany. The author is also sponsored by OKI Electric Industry Co., Ltd, Japan.

(LB), Jumping Window (JW), and Moving Window (MW, or Sliding Window) schemes by using traffic simulation. Finally, we indicate the most suitable policing mechanism based on the above requirements.

2. Requirements for a Policing Function

In CCITT recommendation I. 371 [2] two parameters characterizing the policing performance have been defined.

- Response Time: the time to detect a given non-compliant situation under given reference conditions.
- Accuracy: appropriate control actions on a non-compliant connection and a compliant connection.

There are two types of policing errors which are mis-policing of compliant traffic and the passage of non-compliant traffic. To consider the actuall utility of a UPC function there are some requirements besides the above defined performance parameters. Namely, the policing function should

1) monitor and regulate input traffic for all established VCCs at the User Network Interface,
2) strictly detect violating cells,
3) be transparent for well-behaving cells,
4) be tolerant to cell delay variation,
5) change quickly from non-detecting mode to detecting mode,
6) be implemented economically by hardware, and
7) not depend on services and media.

Requirements 2) and 3) are closely related and characterize the trade-off for policing accuracy of conforming and non-comforming traffic. Requirement 4) is one of the most severe problems of a peak rate policing function.

2. 1 Policing Accuracy

Policing accuracy is an important requirement and is shown in Fig. 1. In this figure arriving cells which are in excess of policing parameter "P," are discareded or tagged. However, the ideal accuracy means both complete detection of violating cells and no-policing of compliant cells, there are two types of errors actually such as overlooking of violating cells and mis-policing of well-behaving cells. These errors are caused by

1) mis-converting from user declared parameters to policing parameters,
2) influence of cell delay variation,
3) mis-declaration of the required bandwidth by the user,
4) asynchronity of traffic parameters between source and network, or
5) asynchronity between detected violation period and actual violation period.

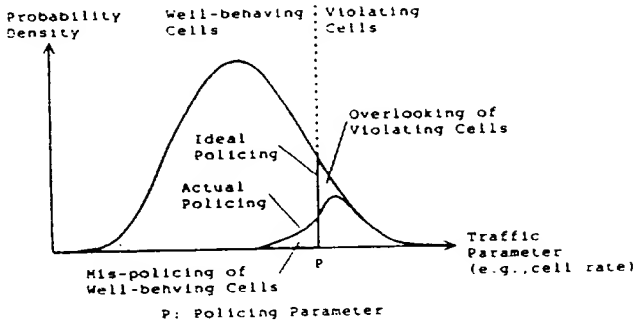


Fig. 1 Policing accuracy.

2. 2 Tolerance to Cell Delay Variation (CDV)

The influence of the CDV on policing performance has been hardly evaluated. In Ref. [12] the influence of the CDV on peak rate policing for a well-behaving connection by using the Leaky Bucket and the Jumping Window mechanisms has been evaluated.

The CDV is caused by cell multiplexing within the CEQ (Customer Equipment) and B-NT (Broadband—Network Termination) which are located between user terminal and the policing devices. As the network congestion is divided into two situations, short-term congestion and long-term congestion [13], the CDV can be also thought as two states, i.e., short-term fluctuations and long-term fluctuations. The former is caused by temporally simultaneous cell arrivals at cell multiplexer and the latter is influenced by a number of long burst arrivals which cause buffer overflow.

In both cases of peak rate and mean rate policing, the CDV should be considered. However, in case of mean rate policing based on a long term average, the influence of the short-term cell fluctuations may be small eough comparing with the policing interval; in case of peak rate policing based on short term average or minimum cell interarrival time, both short-term and long-term cell fluctuations have great influence on the policing performance (see Fig. 2).

Figures 2(a) and (b) show that in case of average rate policing of traffic source which has time fluctuations, the variance of the probability density of the arrival rate is smaller than in case of peak rate policing [3], and even if the CDV is generated, the average rate distribution hardly changes if the measurement period is long enough. However, ideally the policing rate R_P could be set equal to the negotiated rate R_D if the measurement period would be set to the communication period, the UPC mechanism must be tolerant to long-term cell fluctuations because the policing interval time cannot be long enough due to restriction of its hardware implementation and requirement for suitable response time. In case of peak rate policing shown in

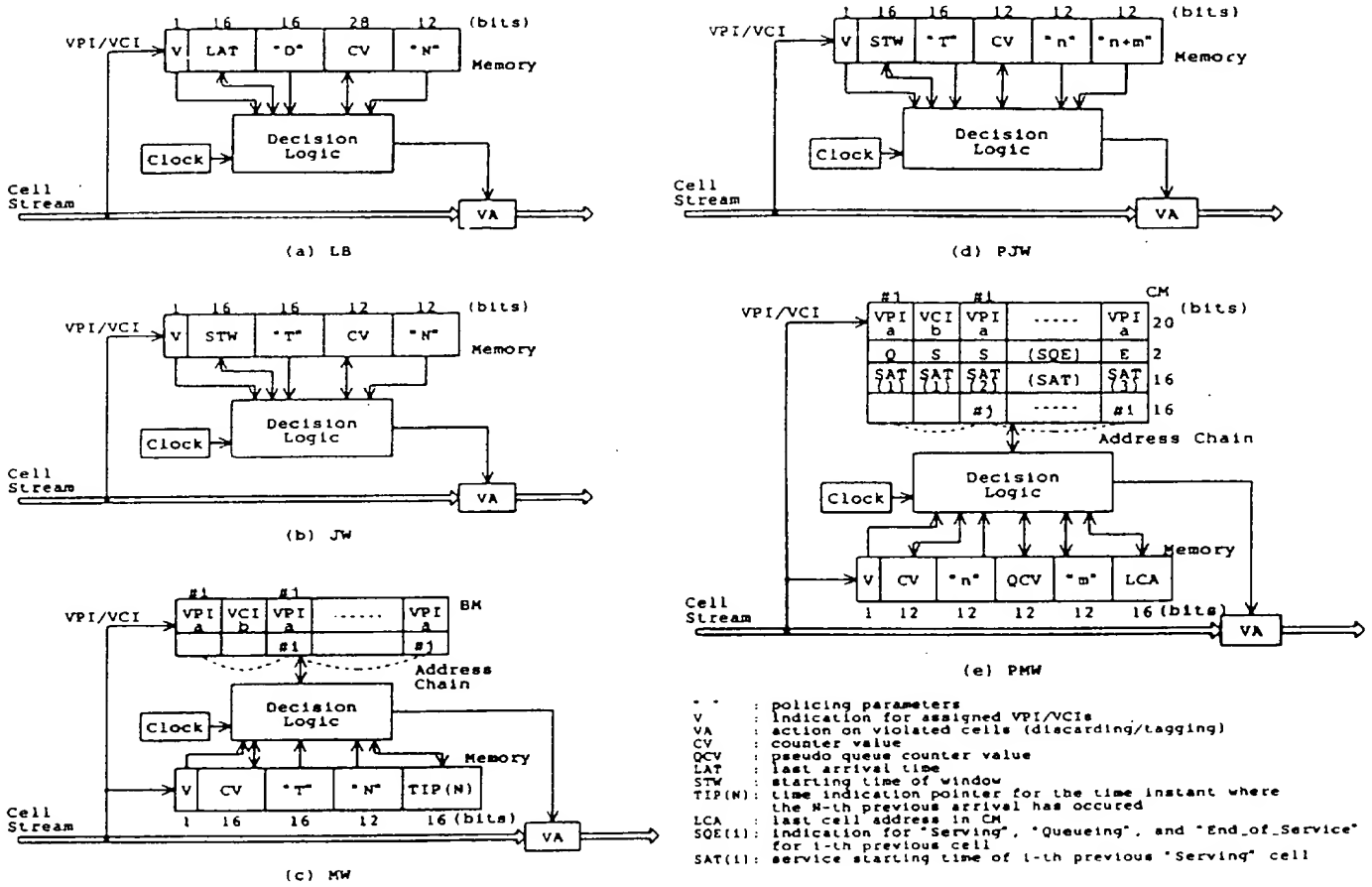


Fig. 4 Implementation examples of policing functions.

places, and the counter limit "N" is equal to "$s+1$") delay loss model represents the LB mechanism exactly, and a stationary analysis has been performed using an iterative method [4], or a fluid flow approach [10]. It should be noticed that cells are not queued in the LB system actually.

Some drawbacks and performance limitations of the LB method are pointed out [8], [9], [15]. In Ref. [15] it is indicated that the LB algorithm can not guarantee the QoS of VCCs for traditional statistical multiplexing and VC/VP bandwidth allocation method. Reference [9] pointed out that the LB method pretends accuracy which does not exist, as the cell arrival process is deformed due to CDV. This is caused by the LB principle which is based on a monitoring of interarrival times of consecutive cells. In Refs. [8] and [9] modified LB algorithms "Jumping Leaky Bucket" and "LB with memory" are introduced.

It is not difficult to implement a LB function as shown in Ref. [16]. Figure 4(a) shows an example configuration of this implementation, where we assumed that

Counter decrementation interval: $D < 2^{16}$

Counter upper limit: $N < 2^{12}$

In the following discussion for the implementations of the policing mechanisms, the operations for overflow of the clock timer are omitted. If we support 2^{20} VP (Virtual Path) and VC (Virtual Channel) connections based on the LB policing, we need about 9.6 MBytes (73 bits $\times 2^{20}$ words) memory, which stores last cell arrival time (LAT), counter value (CV), and two parameters "D" and "N" for each policing VC/VP connection. In a decision logic it is decided wheather an arrived cell violates the negotiated rate or not at each cell arrival by the following steps, where the counter is incremented by "D" at each cell arrival and decremented by 1 at each cell slot, and it needs 28 bits length;

```
at each cell arrival,
    Δt := Current_Time - LAT
    LAT := Current_Time
    CV := max (CV - Δt; 0)
    if CV - D > N x D
        then the cell is discarded or tagged
        else CV := CV + D {the cell is accepted}
```

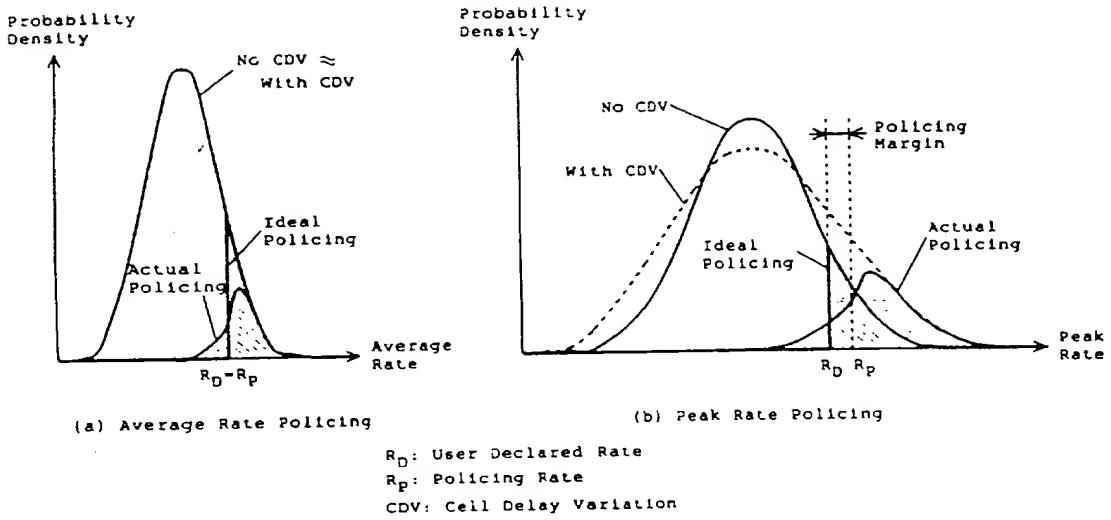


Fig. 2 Influence of the cell delay variation on policing.

Fig. 2(b) it can be supposed that the variance of the rate distribution becomes larger which is indicated by the dotted line if there are multiplexing functions between the source and the policing device, and cells suffer a delay variation. Then if peak rate policing with a user declared rate R_D would be performed, the discarding of well-behaving cells would increase, and the network could not provide the negotiated QoS. Therefore, a safety margin for peak rate policing should be introduced [4], [5], and the policing is performed with a safety rate R_P which is larger than the negotiated rate R_D. Furthermore, the peak rate policing have to be tolerant to both the shor-term and long-term cell fluctuations.

A cell spacing algorithm has been proposed to reduce the impact of cell delay variation by cell queueing [14]. It can be imagined that the spacer consists of a policing and a spacing mechanism with negotiated parameter R_D where the spacing device removes much of the CDV.

2.3 Response Time

The time to detect violating cells which are sent in excess of the negotiated rate should be short to protect the other existing VCCs from the non-compliant source. Therefore, a fast reaction of the UPC function against non-compliant traffic is required. In general, the response time of policing mechanism is shorter as its upper limit of cell counter is smaller.

3. Policing Mechanisms

In this paper we compare the policing mechanisms, Pseudo Jumping Window (PJW), and Pseudo Moving Window (PMW) with existing well-known mechanisms, like the Leaky Bucket (LB), the Jumping

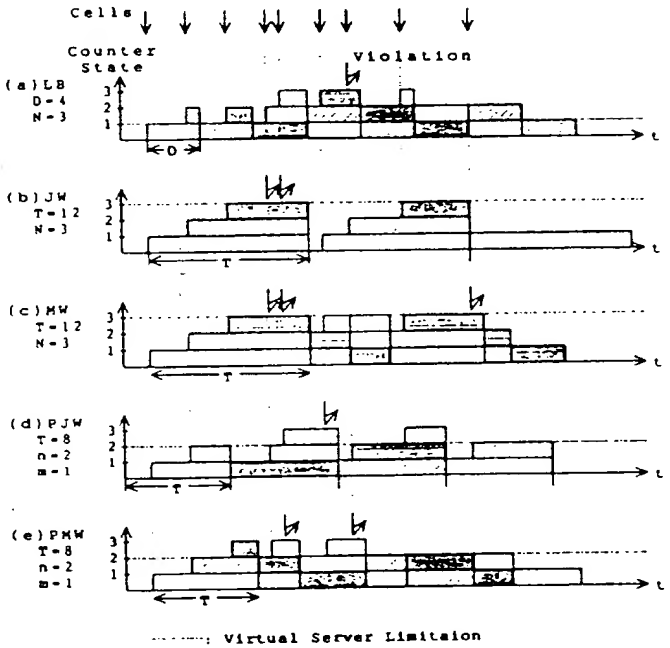


Fig. 3 Example of policing behaviours.

window (JW) and, the Moving Window (MW).

(1) Leaky Bucket (LB)

The LB mechanism can be realized by using an up/down counter which is incremented by 1 at each cell arrival and decremented in fixed intervals "D." The counter value remains in the range from zero to a certain upper limit "N." If the counter value is equal to "N," the policing actions (discarding or tagging) for following cells are taken until the counter value becomes less than "N" again (see Fig. 3(a)). The $G/D/1-s$ (where "s" means the number of queueing

(4) Pseudo Jumping Window (PJW)

As above-mentioned, the existing UPC methods, the LB, JW and MW have some drawbacks with respect to the UPC requirements discussed in Sect. 2. We then propose the PJW and PMW mechanisms as advanced UPC methods.

The PJW is a modified version of the JW and has the same policing parameters "T" and "N" (see Figs. 3(d) and 5(a)). In case of the PJW the parameter "N" is divided into two parameters "n" and "m"($n+m=N$), where "n" means maximum number of cells in each window time (similar to "N" in JW) and "m" means the number of queueing places (credits). In other words PJW is a pseudo (not actually) cell queueing system same as LB and "n" is equivalent to a number of virtual servers. The difference from JW is that even if cells arrived over the allowed number "n" in each window, the cells are not discarded or tagged immediately instead they are "stored" in the queue as long as there are free queueing places. Then after the window ends (i.e., all vitual servers become free), the cells in the queue are assigned into free virtual servers. Obviously when "m" is equal to zero, PJW is equivalent to the JW.

The PJW is a special case of the Jumping Leaky Bucket which has been supposed in Ref. [8]. The Jumping Leaky Bucket has one more policing parameter, which is defined "allowed number of cells per window," and in the case of PJW it is fixed with a value of "n." These two mechanisms, the PJW and the Jumping Leaky Bucket have different approaches which means that the former comes from the JW, and the latter has been derived from the LB mechanism.

The new method PJW can be expected concerning with the follows;

1) it can make the mis-policing probability for long-term fluctuated compliant traffic lower by functioning the pseudo queue same as the LB,
2) it can detect the cell violation more quickly than the JW because it has shorter window interval time as touched upon later, and
3) its shorter window length can also restrict worst cell input with long bursts which causes the network congestion.

Hardware implementation of the PJW can be realized by using the approximation like for the JW implementation. Only extra memory of about 1.6 MBytes (12 bits$\times 2^{20}$ words) is needed under the above conditions (see Fig. 4(d)). And the decision logic performs the following steps to decide whether the cell can be accepted or not;

```
at each cell arrival,
    Δt := Current_Time - STW
    if Δt < T
        then if CV = n - m
                then the cell is discarded or tagged
                else CV := CV + 1 {the cell is accepted}
        else if Δt < K × T {the cell is accepted}
                then while STW < Current_Time - T
                        repeat STW := STW + T
                               CV := max(CV - n, 0)
                     CV := CV + 1
                else STW := Current_Time
                     CV := 1
    endif
```

(5) Pseudo Moving Window (PMW)

The PMW mechanism which is based on the MW can be considered similar to the PJW. PMW is an algorithm adding a pseudo queueing function to the MW and is characterized by the parameters "T," "n" and "m" shown in Figs. 3(e) and 5(b). The $G/D/n-m$ delay loss system is an exact model of the PMW mechanism.

Similar to the PJW mechanism, the PMW can improve the policing performance of the MW with respect to the mis-policing behaviour, the response time and restrictness to the long burst input.

The PMW function cannot be implemented by BM method, because service starting time instants for pseudo queued cells are different from the cell arrival time instants. Then it is required to store the information of cell arrival times and the service starting times for pseudo queued cells and served cells, respectively. Figure 4(e) shows an example of the PMW implementation, where 8.5 MBytes (65 bits$\times 2^{20}$ words) memory and 360 KBytes (44 bits$\times 2^{16}$ words) CM (Common Memory), which stores the above time information, are required. The decision logic is implemented in principle by the following operations. The parameters "n" and "m" are restricted to an upper limit by speed limitation of the hardware components.

```
at each cell arrival,
    X := CV + QCV
    Y := QCV
    while X > 0 and SQE(X) = "Serving" and Current_Time ≥ SAT(X) + T
        repeat SQE(X) := "End_of_Service"
               CV := CV - 1
               if Y > 0
                   then QCV := QCV - 1
                        CV := CV + 1
                        SQE(Y) := "Serving"
                        SAT(Y) := SAT(X) + T
                        Y := Y - 1
               X := X - 1
    if CV = n
        then if QCV = m
                then the cell is discarded or tagged
                else QCV := QCV + 1
                     store VPI/VCI value in CM within the address chain
                     SQE(I) := "Queueing"
        else CV := CV + 1
             store VPI/VCI value in CM within the address chain
             SQE(I) := "Serving"
             SAT(I) := Current_Time
    discharge all "End_of_Service" cells from CM
    update the address chain in CM
    update LCA
```

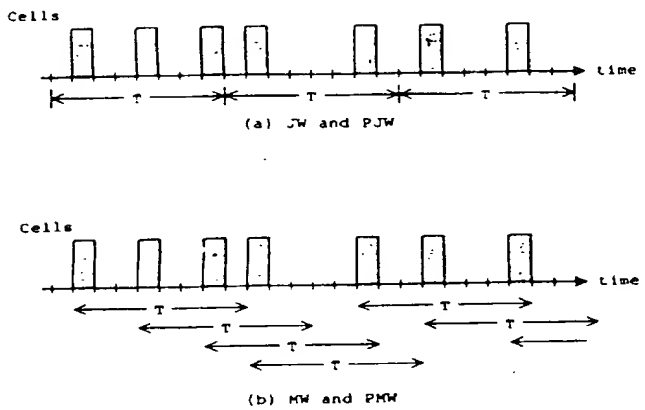


Fig. 5 Principle of the policing mechanisms ($T=8$).

(2) Jumping Window (JW)

JW mechanism can be defined as a counting algorithm that counts arriving cells in a fixed time interval (window) "T". "N" cells are allowed in each window "T" and a new window starts immediately when the preceding window ends (see Figs. 3(b) and 5(a)). The counter value is reset to zero at the beginning of each window. If the conter value is equal to "N," following cells are discarded or tagged until the window terminates. The stationary performance of the JW is evaluated in Ref. [4] by using the counting process of the cell arrivals.

In Ref. [4], it is pointed out that the JW mechanism cannot guarantee the required QoS (e.g., cell loss probability) because of its policing strictness and since it needs large counter limit, the JW cannot detect the violation quickly. It should be noticed that window-based UPC method such as the JW cannot tolerate the longer-term cell fluctuations than its policing interval time.

Implementation complexity of the JW function is similar to the LB as can be seen in Fig. 4(b). We assumed the following counter limits for the JW, MW, and PJW mechanisms:

Window size: $T<2^{16}$

Maximum number of cells in each window:

$$N<2^{12}$$

Under this assumption about 7.5 MBytes (57 bits$\times 2^{20}$ words) memory is necessary to implement the JW function. Then the decision logic in the figure operates according to next steps, where to avoid a use of a divider the STW (starting time of window) is reset with the current time, if there is no cell arrival for a time interval $K\times T$ (K is a constant value of integer):

```
at each cell arrival.
    Δt := Current_Time - STW
    if Δt < T
        then if CV = N
                then the cell is discarded or tagged
                else CV := CV + 1 {the cell is accepted}
        else CV := 1 {the cell is accepted}
            if Δt < K x T
                then while STW < Current_Time - T
                        repeat STW := STW + T
                else STW := Current_Time
    endif
```

(3) Moving Window (MW)

The MW mechanism is also called Sliding Window (SW) or Dangerous Bridge (DB) and has the same parameters "T" and "N" as the JW mechanism. The difference is that a separate window starts at every cell slot or at least at every cell arrival (see Figs. 3(c) and 5(b)). The MW scheme is exactly represented by a $G/D/N$ loss model, and if a Poisson process is used for the cell arrivals, the violation probability of the MW mechanism can be easily calculated by using Erlang's loss formula.

Similar to the JW mechanism, the MW has significant drawbacks of high mis-policing probability for compliant traffic which suffers long-term fluctuations, and slow response time to detect the violation due to its large counter limit [4].

The hardware implementation of the MW function is not easy and needs some contrivances. Refs. [16], [17] and [18] show the implementation methods based on a BM (Bridge Memory) which is a FIFO (First In First Out) type memory for common use among the all VP/VC connections. In case of BM-P (BM-Pointer) method, which has been proposed in Ref. [17], VPI/VCI values of the arrived cells are stored in the BM within an address chain. In Fig. 4(c) an example of a MW realization using the BM-P method is illustrated. According to the above assumptions 8 MBytes (61 bits$\times 2^{20}$ words) memory and 300 KBytes (36 bits$\times 2^{16}$ words) memory are needed to implement the MW function to support 2^{20} VP and VC connections. And the decision logic, which is relatively complex compared with the LB and the JW mechanisms, is based on following operations:

```
at each cell slot.
    if a cell arrival
        then if CV ≥ N
                then Δt := Current_Time - TIP(N)
                    if Δt < T
                        then the cell is discarded or tagged
                        else CV := CV + 1 {the cell is accepted}
                else CV := CV + 1 {the cell is accepted}
            if the cell is accepted
                then store VPI/VCI value in BM within the address chain
                    if CV ≥ N
                        then update TIP(N)
                else store "empty cell" in BM
        else store "empty cell" in BM
    endif
    discharge first cell from BM
    if the discharged cell is not "empty"
        then CV := CV - 1
```

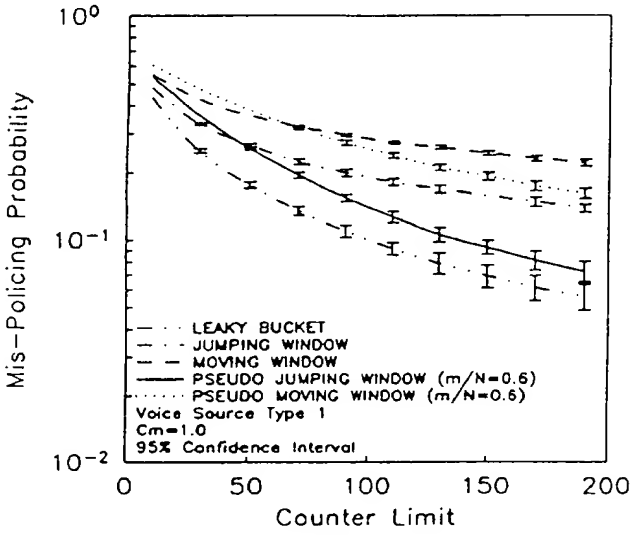


Fig. 7 Influence of the counter limit with $C_m = 1.0$.

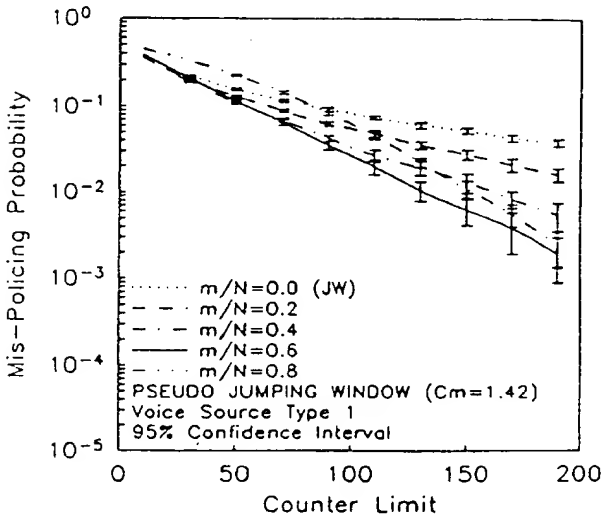


Fig. 9 Effectiveness of the pseudo queueing function for the PJW.

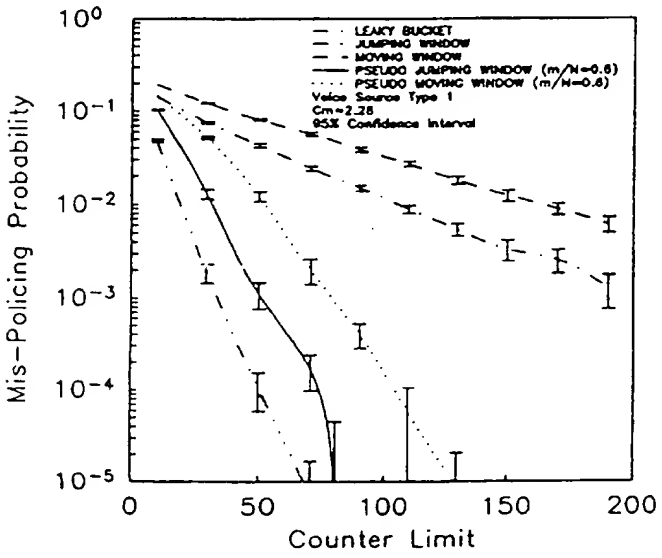


Fig. 8 Influence of the counter limit with $C_m = 2.28$.

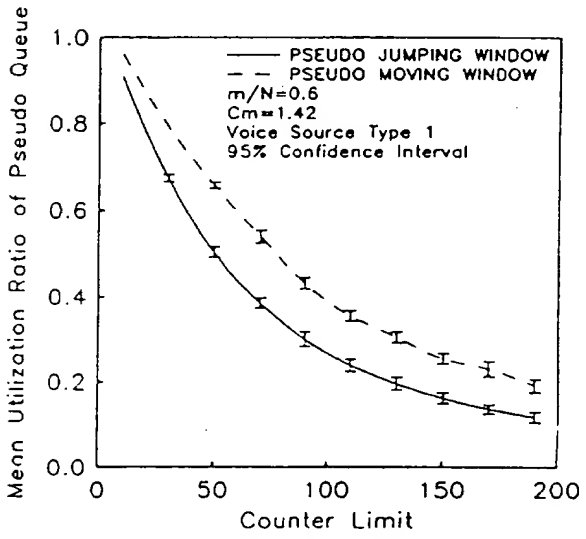


Fig. 10 Utilization ratio of the pseudo queue.

satisfactory responsiveness of the mechanisms.

4.2 Effectiveness of Pseudo Queueing

The effectiveness evaluation of a pseudo queueing function for the PJW algorithm is shown in Fig. 9. Comparing with $m/N = 0.0$ (JW) case, the mis-policing probability becomes lower when m/N is 0.2, 0.4, and 0.6, and it indicates that a pseudo queueing function leads to a lower mis-policing probability for well-behaving sources. When $m/N = 0.8$, the number of allowed cells per window (i.e., the number of servers) is so small that the mis-policing probability increases for small counter limits. For the following simulations we use a value of 0.6 for an m/N factor of the PJW and PMW.

In Fig. 10 the mean utilization ratio of the pseudo queue for the PJW and PMW mechanisms, where $m/N = 0.6$ and $C_m = 1.42$, is shown. It is obvious that the pseudo queueing function takes more effect as the counter limit is smaller.

4.3 Policing Accuracy

The policing performance against violated cells caused by a long-term fluctuation of the traffic or wrong declaration by the user is shown in Fig. 11. The normalized average cell rate denotes the ratio of actual

4. Performance Analysis of the Mechanisms

Figure 3 shows the behaviour of the mechanisms for mean rate policing with an average cell rate $R=1/4$ times the line capacity. In this figure we assumed $N=3$ cells and $T=12$ cell units for the JW and MW mechanisms, $N=3$ and $D=4$ for the LB mechanism, and $n=2$, $m=1$ and $T=8$ for the PJW and PMW mechanisms. This figure indicates that there are 9 cell arrivals in 32 cell units time interval and the cells suffer long-term fluctuations beyond the policing interval time. It is obvious that the LB can tolerate the long-term cell fluctuations but not detect a violating cell quickly and it detects a well-behaving cell possibly because pseudo queueing method such as the LB cannot discriminate non-compliant cells from compliant cells easily. On the other hand, the JW and MW, however, can detect the violation quickly, they have the strict behaviour and therefore the mis-policings because window-based mechanisms cannot tolerate the long-term fluctuations. The PJW and PMW we proposed in last section are the composite methods of both pseudo queueing function and window-based algorithm to detect the violating cells quickly and to tolerate the long-term cell fluctuations.

Similar to Ref. [4] we introduce a policing margin factor (overdimensioning factor) C_m, which is defined as follows,

$$C_m = R_p/R = \begin{cases} (1/D)/R, & \text{for LB} \\ (N/T)/R, & \text{for JW and MW} \\ (n/T)/R, & \text{for PJW and PMW} \end{cases} \quad (1)$$

where R_p is the average policing rate and R is the negotiated average rate. The parameters D and T are determined by using Eq. (1). By introducing $C_m(>1)$ it is possible to accomodate the CDV. It should be noticed that increasing C_m results in the deterioration of the ability to detect non-compliant cells, and an increasing counter limit "N" causes in increase of the response time. Also note that if the counter limits, N of the JW and MW, $(n+m)$ of the PJW and PMW, are set to same value, the window length Ts of the PJW and PMW are shorter than those of the JW and MW because of $n<N$.

To simulate the policing mechanisms we used a two-phase burst/silence source model to represent the statistical fluctuations of the cell arrival process. In this model which has two states, burst and silence, only during burst phase cells arrive periodically with inter-arrival time d. The number of cells within a burst is distributed according to a geometric distribution with mean $E[X]$ and the silence period is distributed according to a negative-exponential distribution with mean $E[S]$ as illustrated in Fig. 6. As shown in Table 1 we applied this model to two types of packetized voice where it is assumed that both the bandwidth and the silence period are compressed using 32 kb/s ADPCM and speech detection, respectively, and 48 Bytes information field per cell. Both packetized voice sources have the same peak rate and the same average rate. By using such source model, we evaluated the performance of the UPC mechanisms as a case study.

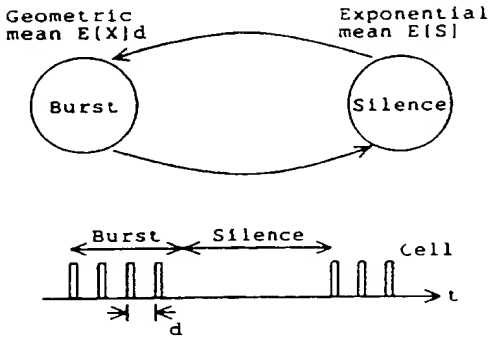


Fig. 6 2-phase burst/silence model.

Table 1 Parameters for the input traffic sources.

Source Type	E[X]	E[S]	d	R	Ave.Rate	Peak Rate
Voice Packet 1	29.3	650.0 ms	12ms	29.3cell/s	11.2kb/s	32.0kb/s
Voice Packet 2	7.0	155.29ms	12ms	29.3cell/s	11.2kb/s	32.0kb/s

4.1 Influence of the Counter Limit

Figures 7 and 8 show the influence of the counter limit on the mis-policing probability for mean rate policing with a policing margin factor $C_m=1.0$ and $C_m=2.28$, respectively, using voice packet 1 which can be thought as a long-term fluctuated compliant source. A parameter called m/N value (ratio of the number of queueing places and the counter limit) is set to 0.6 for the PJW and PMW mechanisms. The mis-policing probability of the UPC function for compliant traffic should be as possible as small. Generally it is said that in an ATM network the cell loss probability for end-to-end transfer should be in the order of 10^{-9} (and 10^{-10} per node) to ensure the QoS of well-behaving sources. All mechanisms with $C_m=1.0$ shown in Fig. 7 would need counter limits more than 10^8, to satisfy the above loss probability criterion, if cells are discarded by the policing function. With $C_m=2.28$ as shown in Fig. 8, the LB, PJW, and PMW, which have the pseudo queueing function and tolerate to the long-term cell fluctuations, can achieve the above mis-policing probability with a counter limit of about 2×10^2, which can be realized by an 8 bit counter. On the other hand, for the JW and MW mechanisms, which have no pseudo queueing function, the safety margin C_m has to be close to the ratio of peak to average cell rate to achieve a

performance for the compliant voice source can be neglected mostly. For the JW mechanism introducing of the approximation is possible similar to the PJW mechanism. On the other hand, as shown in Fig. 14 evenif the same approximation for the LB implementation would be applied to reduce the memory size (in this case CV needs only 12 bits length), the large value of K, more than 20, would be required and it would make the implementation complex. The reason is that in case of the JW and the PJW K means $K \times T$ time interval where no cell arrives, and in case of the LB mechanism K determines $K \times D$ time interval where no counter decrementation occurs, and it is obvious from Eq. (1) that $D = T/N$(JW) or T/n(PJW).

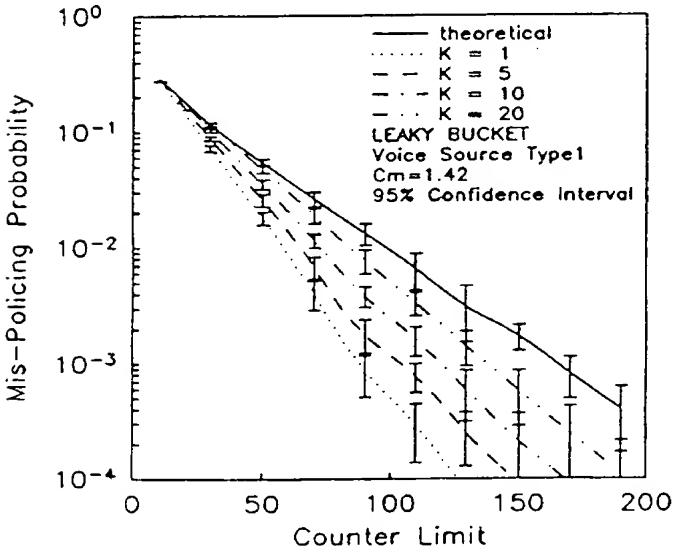


Fig. 14 Performance of the LB based on approximated implementation.

5. Overall Comparison

In Table 2 we compare all policing mechanisms which have been considered within this paper, concerning policing strictness, accuracy for non-compliant traffic, response time, implementation complexity which includes memory size as well as decision logic complexity, and tolerance to the short-term CDV. It indicates that all mechanisms have some drawbacks, and for the case of mean rate policing the PJW and the LB are the most suitable mechanisms because these two methods have only shortcomings of the required memory size for the implementations. On the other hand, the JW and MW have some significant drawbacks on policing performance and the PMW cannot be implemented by hardware easily. Here it should be noticed that the first and second comparing items in Table 2 are requirements of the QoS for each user, while the third and fourth comparing items are concerned with network design such as resource management.

6. Conclusions

For ATM networks it is said that cell level traffic regulation which does not exist in conventional switching systems based on STM (Synchronous Transfer Mode), is indispensable. In this paper we have discussed and evaluated one of those functions, the UPC or policing function. The policing function influences the performance quality of cell multiplexing and switching after the policing devices, and the end-to-end communication quality of users (i.e., the QoS). Therefore, it is necessary to select the most suitable policing mechanisms and parameters through the simulation.

First, we clarified some requirements for policing functions, especially the accuracy for non-compliant source, the traffic transparency for compliant sources,

Table 2 Overall comparison of the mechanisms for mean rate policing.

Comparison Items		LB	JW	MW	PJW	PMW
Realization of required cell loss probability (Tolerance to long-term CDV)		Good	Poor	Poor	Good	Good
Tolerance to short-term CDV		Good	Good	Good	Good	Good
Detection accuracy for non-compliant cells		Fair	Fair	Good	Fair	Good
Response time		Good	Poor	Poor	Good	Good
Implementation	Required memory quantity	Poor	Good	Fair	Poor	Fair
	Logic complexity	Good	Good	Fair	Good	Poor*

* It is required to restrict the parameters "n" and "m" to certain upper limits.

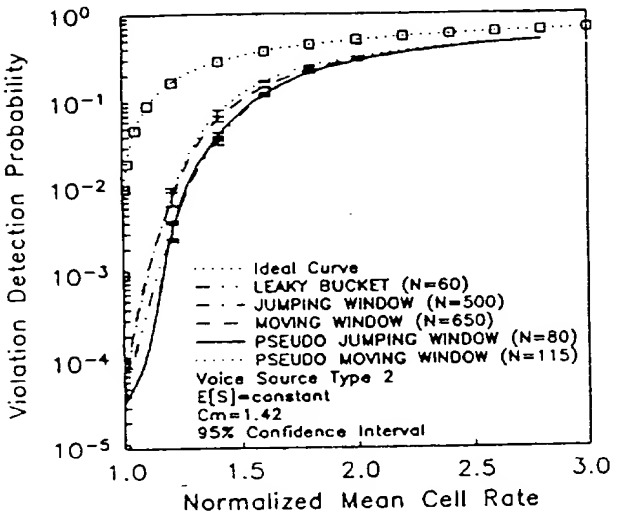


Fig. 11 Performance comparison for a non-compliant source.

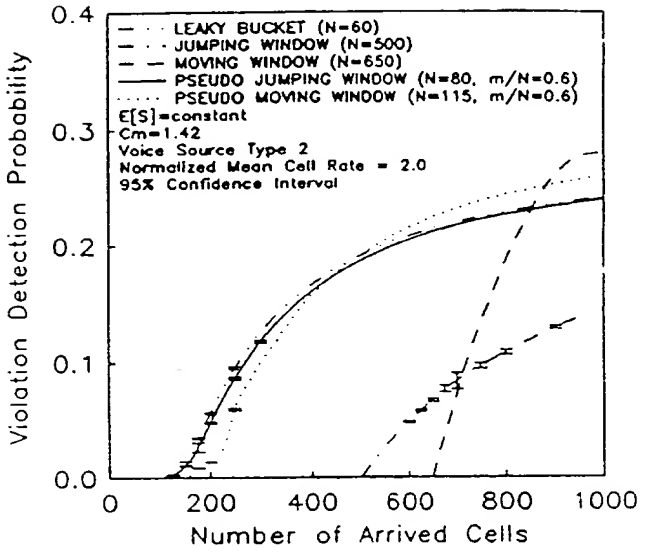


Fig. 12 Transient response to a non-compliant source.

average cell rate, which is changed by increasing burst duration keeping $E[S]$ constant as 147.72 ms, and the negotiated average cell rate. The policing margin factor C_m has been set to 1.42, which is equivalent to the mean value between peak and average rate. In order to compare the mechanisms, the counter limits are set to the values shown in the figure, and all mechanisms have almost the same violation detection probability of about 3×10^{-5} at the nominal average cell rate. As shown in the figure the MW and PMW mechanisms achieve the best results. The other mechanisms—LB, JW and, PJW—show almost the same performance. For the case $E[X]=5.0$, we obtained similar results.

4.4 Response Time—Dynamic Behaviors

The response time to source violation depends largely on the counter limit value. In Fig. 12 the dynamic behavior of the policing mechanisms is illustrated for a non-compliant source for which the normalized average cell rate is set to 2.0 by choosing $E[X]=21.8$ and $E[S]=147.72$ ms, and the counter limits are set to the values indicated in Fig. 11, is illustrated. The violation detection probability of each mechanism increases after the first virtual server termination which means either the first counter decrementation for the LB or the first window termination for the window based mechanisms. The violation detection probability of the MW rises rapidly but not quickly. LB, PJW, and PMW have almost the same dynamic behavior. If all mechanisms would have the same counter limit value, MW would be the most desirable with respect to the response time. In actual applications LB, PJW, and PMW are most suitable, because MW and JW need large counter limits to realize the required mis-policing probability of 10^{-9} as shown in Figs. 7 and 8.

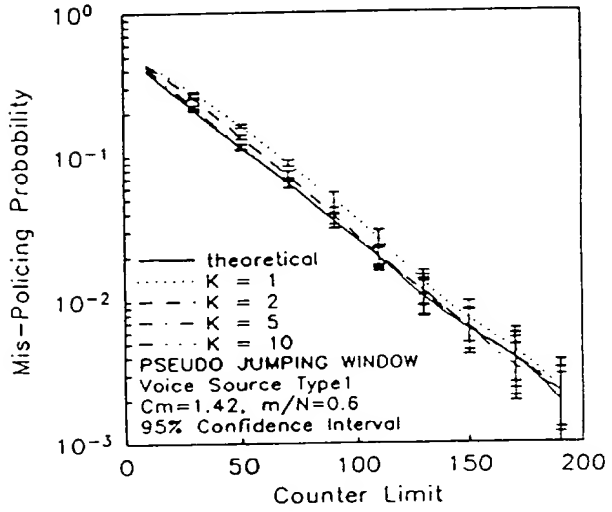


Fig. 13 Performance of the PJW based on hardware implementation.

4.5 Policing Performance Based on Hardware Implementation

To realize the policing functions easily the approximations of hardware implementations are taken for the JW and the PJW mechanisms. In Fig. 13 the comparison of the PJW policing performance based on the implementation with theoretical performance is illustrated. As shown in the figure, if the constant value K (K means that the window starting time STW is set to current time when there are no cell arrivals in time interval $K\times T$) is set to more than 5, the difference of the hardware implementation from the theoretical

and the tolerance for CDV. Next, two policing mechanisms PJW and PMW were proposed based on pseudo cell queueing which are more flexible than JW and MW, respectively, from view points of the policing accuracy for long-term fluctuated compliant traffic, fast response ability, and the strictness to the cell traffic with long bursts. Then we have assessed the performance of the above mechanisms and well-known LB, JW, and MW mechanisms by simulation. The main issues we have studied comprise the influence of the counter limit, effectiveness of pseudo queueing functions, accuracy, transparency, response time (dynamic behavior), and hardware implementation. By performance comparison we also indicated that for the JW and the PJW mechanisms the approximation can be applied to implement the functions easily. Finally, we compared those five mechanisms based on policing performance ability, implementation complexity and tolerance for CDV, and the results show that the PJW and the LB are the most suitable and desirable mechanisms for mean rate policing, because the other three mechanisms have some significant drawbacks on their policing performance and implementation to monitor and regulate source traffic in ATM networks.

Acknowledgement

The author would like to acknowledge Professor Dr.-Ing. Paul J. Kuehn and Mr. Hans Kroener who gave useful advice, and Dr. Erwin P. Rathgeb, Siemens AG, Muenchen, F. R. Germany, who offered his simulation programs readily.

References

[1] *CCITT Study Group XVIII I. Series Draft Recommendations*, Geneva, 1991.

[2] *CCITT Study Group XVIII Recommendation I. 371*, "Traffic Control and Congestion Control in B-ISDN," Geneva, 1992.

[3] Hirano, M., "Policing Parameters for Variable Bit Rate Traffic," *IEICE Technical Report*, SSE90-92, 1990.

[4] Rathgeb, E. P., "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 3, pp. 325-334, Apr. 1991.

[5] Rathgeb, E. P. and Theimer, T. H., "The Policing Function in ATM Networks," *Proc. of International Switching Symposium (ISS) '90*, Stockholm, pp. 127-130, May 1990.

[6] Hemmer, H. and Huth, P. T., "Evaluation of Policing Function in ATM Networks," *Proc. of Workshop on Queueing, Performance, and Control in ATM, 13th International Teletraffic Congress (ITC13)*, Copenhagen, pp. 111-116, 1991.

[7] Monterio, J. A. S., Gerla, M. and Fratta, L., "Input Rate Control for ATM Networks," *Proc. of Workshop on Queueing, Performance, and Control in ATM, 13th International Teletraffic Congress (ITC13)*, Copenhagen, pp. 117-122, 1991.

[8] Iversen, V. B. and Nielsen, A. B., "Traffic Management in ATM Networks Based on a Counting Mechanism," *Proc. of Workshop on Broadband Communications*, Estoril, pp. 281-291, Jan. 1992.

[9] Lague, B., Rosenberg, C. and Guillemin, F., "A Leaky Bucket with Memory," *Proc. of Workshop on Broadband Communications*, Estoril, pp. 256-266, Jan. 1992.

[10] Butto, M., Cavallero, E. and Tonietti, A., "Effectiveness of the "Leaky Bucket" Policing Mechanism in ATM Networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 3, pp. 335-342, Apr. 1991.

[11] Yin, N. and Hluchyj, M., "Analysis of the Leaky Bucket Algorithm for On-Off Data Sources," *Proc. of GLOBECOM '91*, Phoenix, pp. 254-260, 1991.

[12] Roberts, J. and Guillemin, F., "Jitter in ATM Networks and its Impact on Peak Rate Enforcement," *Performance Evaluation*, vol. 16, no. 1-3, pp. 35-48, North-Holland, Nov. 1992.

[13] Kroener, H., Theimer, T., Briem, U., "Queueing Models for ATM Systems—A Comparison," *Proc. of International Teletraffic Congress (ITC), 7th Specialist Seminar*, Morristown, Paper No. 9.1, Oct. 1990.

[14] Guillemin, F., Boyer, P. and Romoeuf, L., "The Spacer-Controller: Architecture and First Assesments," *Workshop on Broadband Communications*, Estoril, pp. 294-304, Jan. 1992.

[15] Yamanaka, N., Sato, Y. and Sato, K., "Performance Limitation of Leaky Bucket Algorithm for Usage Parameter Control and Bandwidth Allocation Methods," *IEICE Trans. Commun.*, vol. E75-B, no. 2, pp. 82-86, Feb. 1992.

[16] Dittman, L., Jacobsen, S. B. and Moth, K., "Flow Enforcement Algorithms for ATM Networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 3, pp. 343-350, Apr. 1991.

[17] Kusayanagi, M., Takeo, H, Ogura, T., Iguchi, K. and Yamaguchi, K., "Policing Techniques for ATM Subscriber Circuit," *1992 Spring Natl. Conv. Rec. IEICE*, B-681.

[18] Nishihara, M., Kurano, T. and Akashi, F., "A Consideration on Policing Techniques in ATM Networks," *1992 Spring Natl. Conv. Rec. IEICE*, B-682.

Kiyoshi Shimokoshi received the B.E. degree from the University of Tokyo, Japan, in 1987. In 1987, he joined OKI Electric Industry Co., Ltd., Tokyo, Japan, where he has been engaged in research and development on ATM switching systems. Since 1991 he is a visiting researcher at Institute of Communications Switching and Data Technics, the University of Stuttgart, F. R. Germany, and his current research interests include performance analysis and traffic control technology for B-ISDN.

p. 1224-1233

Study Of A Two-Level Flow Control Scheme and Buffering Strategies

Fengmin Gong

Information Technologies Division
MCNC
RTP, NC 27709

Gurudatta Parulkar

Department of Computer Science
Washington University
St. Louis, MO 63130

Abstract

This paper proposes using an explicit rate control together with a simple window control for flow control in high-speed networks with resource reservations. The design justifications, the evaluation study, and the implementation results are presented. In particular, the idea of hard ACK *and soft* ACK *in the context of end-to-end flow control is explored through simulation and the results are discussed.*

1 Introduction

Many flow control schemes have been proposed over the years. Doeringer and his colleagues have done an extensive survey of these schemes [6]. For example, a sliding window scheme is used in the TCP (Transmission Control Protocol) [8] and rate control schemes are used in VMTP [2] and NETBLT [4]; SNR also makes use of a window-based flow control but it uses periodic state exchange to reduce control latency [15]; XTP supports window-based and rate-based flow controls as two options [3]. However, most of these schemes still assume network level support with best-effort services and have not paid sufficient attention to emerging applications that require service guarantees (e.g., network distributed computing and visualization). We propose a flow control scheme that takes into account the requirements of the new applications and takes advantage of the reservation-based networking environment (e.g., Asynchronous Transfer Mode (ATM) networks). The design, evaluation, and implementation of the scheme are the main subjects of this paper.

The rest of this paper is organized as follows. Section 2 describes the two-level flow control scheme. Section 3 examines several important performance questions regarding the end-to-end flow control and presents the analysis and simulation results. Some implementation results are presented in Section 4. Finally, a summary is presented in Section 5.

2 Proposed solution

There are three main principles that guided our design:

- Direct exchange of end-to-end control messages should be more effective than relying on the "back-pressure" to build up hop by hop for end-to-end speed matching between the two end applications.
- Congestion is affected by a wide range of control mechanisms and policies at the data link level, the network level, and the transport level [12]. A successful congestion control solution is possible only if systematic design decisions are made at all these levels. Therefore, considerations should be made to help the congestion control in underlying networks when designing an end-to-end flow control mechanism. However, adverse interactions between congestion control, flow control, and error control should be avoided [4].
- Control data granularity refers to the smallest unit of data that the flow control mechanism acts upon. Possible granularity levels include the traditional byte, packet, and segment which defines the smallest data unit that an application wishes to access independently. Choice of granularity should be determined from the tradeoff between the effectiveness of control and the control overhead.

2.1 Overall structure

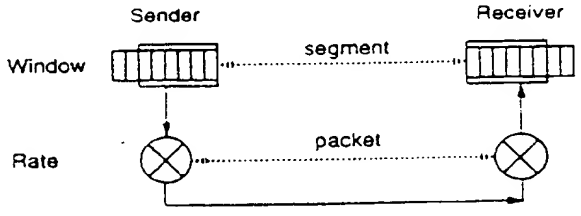


Figure 1: Two-Level Flow Control Structure

Figure 1 illustrates flow control mechanisms as they reside in a sender and a receiver. The data flows from the sender to the receiver. At the lower level there is a rate control mechanism that regulates the data flow into the underlying network according to a rate specification. The rate is determined at the connection setup

with the underlying network and will not be modified for the duration of the connection. At a higher level, a simple window mechanism resolves whether a data segment is eligible for transmission through the rate control. As indicated by the dashed lines, rate control is on a packet basis while the window control uses segment granularity. While we believe that the rate control function is necessary at the transport level, the rate control mechanism may be combined with one at a lower level to avoid duplication.

There are several advantages from this two-level structure. First, when the rate control is used in concert with a window control it helps to decouple the flow control from the congestion control. The rate control ensures that the traffic generated by the local application conforms to a negotiated rate specification. Thus, the traffic into the underlying network is more predictable so that the network can deploy more efficient congestion control strategies. The window mechanism allows a receiver to control the speed at which the sender is transmitting data segments, according to the receiver's speed and buffer availability. Second, two levels of granularity can be used for effective and efficient control. The rate control is at the packet level which ensures a finer granularity of control for data flowing into the network. The window control uses segment granularity. The larger segment granularity reduces the overall control overhead. Third, adverse interactions between error and flow controls are reduced. We use special control messages for window advancement. These messages are different from the acknowledgments for reception of data segments. Therefore, even when the receiver needs to slow down the sender, it can still continue to send acknowledgments as necessary so that the sender will not generate false retransmission by confusing a "slow-down" request as a packet loss timeout, which is the case with TCP.

It should be noted that this flow control scheme is part of an effort to provide efficient interprocess communication (IPC) support for networked pipelined applications. In the effort, a segment streaming IPC was proposed, and a segment streaming transport protocol (SSTP) was designed and implemented. In particular, an application-oriented error control scheme has been designed to work with this flow control scheme [10]. Details of the entire work are available in [11].

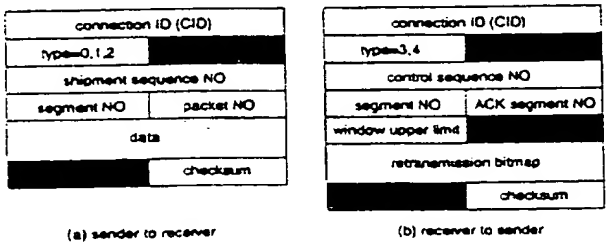


Figure 2: Error Control Packet Formats

2.2 Packet formats

There are five types of packets used for data transfer and flow control. We assume that a connection has already been established between the two transport entities so we do not need to deal with connection management packets here. There are two basic packet formats as shown in Figure 2. The width of each row is assumed to be 4 bytes.

Each packet starts with a 4-byte connection identifier field (CID) followed by a 2-byte type field and ends with a 2-byte checksum field. The CID identifies all the packets belonging to a connection. The type field contains a unique code number for each packet type. Of the five types, type 0, 1, and 2 are for conveying information from sender to receiver; type 3 and 4 are used in the reverse direction. The checksum field is filled by the sender and then used by the receiver to detect packet corruption. The rest of the fields are summarized as follows:

Data Packet (type 0): Data packet is for carrying application data from the sender to the receiver. The shipment sequence number indicates the sequence at which the packet was last shipped from the sender and is used for loss detection. The segment number along with packet number identifies the position of this data block in the application's data stream. The data itself is carried in the data field.

Keep-Alive Packet (type 1): Keep-alive packet is for informing the receiver that the sender is at a pause. Periodic transmission of this packet serves two purposes: (1) to prevent the receiver from closing the connection and (2) to provide the context (shipment sequence number, segment number, and packet number) for the receiver to detect losses in earlier segment transmissions. This format is the same as that of the data packet except that it does not have a data field.

EndStream Packet (type 2): End-of-stream packet is for informing the receiver of the end of a segment stream. It also allows the receiver to make prompt detection of losses occurring in the last portion of the segment stream. This format is the same as that of the keep-alive packet.

PACK Packet (type 3): A PACK (positive ACK) packet serves three purposes: (1) to acknowledge the acceptance of the segment specified by segment NO, (2) to cumulatively acknowledge all the segments with segment number below ACK segment NO, and (3) to optionally advance the sender's window by specifying a new window upper limit. The control sequence NO is used by the sender to detect and discard duplicate PACK packets. Note that control messages for window advancement is in a field separate from the acknowledgment. Therefore, even when the receiver needs to slow down the sender, it can still continue to send acknowledgments as necessary without window advancement so that the sender will not gen-

as the window? These questions have been explored using simple analyses and simulations.

We have used two performance measures, *throughput* and *delay*. To concentrate on the performance of the protocol mechanism itself, the maximum achievable throughput when application process is not a bottleneck is considered. Furthermore, a throughput normalized against the connection rate is used in order to obtain a direct measure of how efficient the mechanism can utilize the connection. Specifically, *Throughput Efficiency* is defined as *the ratio between the ideal segment transmitting time (segment size divided by the connection rate) and the actual time required on average to successfully deliver one segment.*

The *End-to-End Delay* of a segment is defined as *the time elapsed from the start of transmission at the sender till the successful acceptance of the segment at the receiver*. This delay definition is from the receiver's point of view. If it were defined from the sender's point of view, there will be additional time for an acknowledgment to reach the sender. The former definition is chosen because it reflects exactly when the segment is available for consumption at the receiver.

Following the basic definition of end-to-end delay, *Average End-to-End Delay* and *Maximum End-to-End Delay* are defined respectively as the average and the maximum of end-to-end delays for a given number of segments.

Delay analysis

The main purpose of this analysis as well as the one for throughput to be presented later are: (1) to gain insights through the derivation of simple performance expressions and (2) to verify the simulation results.

We approximate the average end-to-end delay for transmission of multiple segments by the average delay of a single segment transmitted in isolation. Let the packet size be l bytes, the segment size be s packets, and the data rate be A bps (bits/second). The packet transmitting time is $t_p = (8 \times l)/A$ seconds. The following specific assumptions apply to this analysis:

- A packet is lost if it is either corrupted or dropped in the underlying network. Each new packet transmission can be lost independently with probability p.
- An acknowledgment (PACK or SNAK) always comes back to the sender t_a time after the transmission of the last packet of a packet group[1]. Acknowledgment delay t_a is determined as $t_a = RTD + 3 \times t_p$. Here we assume that the receiver will wait until $3 \times t_p$ after the reception of the segment to send an advancement message and that other processing overhead at the receiver is negligible. The number 3 is somewhat arbitrary but it signifies the fact that there will be a delay equivalent to several t_p for the receiver to cope with the out-of-sequence packet delivery in the network and detect packet losses.
- Application requires 100% reliable delivery of all segments.

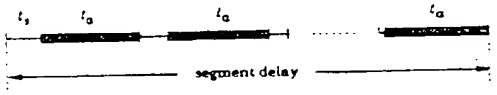


Figure 4: A General Segment Delivery Scenario

Figure 4 depicts a general scenario for the delivery of a segment. A segment is successfully delivered after at least one round of attempt. The first round consists of the transmitting time t_s for the whole segment followed by the waiting time t_a for the acknowledgment (the first shaded area in the figure); in each subsequent round, the lost packets from the previous round are retransmitted and another t_a has to pass by before an acknowledgment is received; no more attempt will be needed when a positive acknowledgment is received. Let D_s denote the average segment delay as shown in the figure, it contains clearly three parts:

$$D_s = t_{transmit} + t_{retransmit} + t_{wait}$$

The segment transmitting time is simply $s \times t_p$. The second term, the mean retransmission time for the segment is $s \times t_p \sum_{i=0}^{\infty} i \times p^i \times (1-p)$. As each round has a constant waiting time t_a, the mean waiting time equals t_a times the mean number of rounds, which is $t_a \sum_{i=1}^{\infty} i \times [(1-p^i)^s - (1-p^{i-1})^s]$. Therefore:

$$\begin{aligned} D_s &= st_p + st_p \sum_{i=0}^{\infty} ip^i(1-p) \\ &\quad + t_a \sum_{i=1}^{\infty} i[(1-p^i)^s - (1-p^{i-1})^s] \\ &= \frac{t_s}{1-p} + t_a \sum_{i=0}^{\infty} [1-(1-p^i)^s] \end{aligned}$$

Derivation of D_s assumed sender's point of view. Let D_r be the average delay as defined from the receiver's view point, it is then related to D_s as follows:

$$D_r = D_s - \frac{1}{2} RTD - 3t_p$$

It should be mentioned that a similar analysis has been presented in [14] for a selective acknowledgment scheme that uses packet granularity (i.e., there is no concept of a segment). In that analysis, expressions were derived for both average delay (mean) and the variance, while the main interest here is the mean value which allows a simpler derivation.

[1] A packet group refers to all the packets associated with a segment. For example, a segment is itself a packet group when it is first transmitted. During a retransmission all packets to be retransmitted for a segment define a new packet group.

erate false retransmission by confusing a "slow-down" request as a packet loss timeout.

SNAK Packet (type 4): A SNAK (selective negative ACK) packet is used to request retransmission of the packets specified by the retransmission bitmap and to cumulatively acknowledge all those segments with segment number below ACK segment NO. The most recently granted window limit is contained in field window upper limit to protect against the loss of the last window advancement message. The difference between this packet and the PACK packet is that it has a retransmission bitmap. The bitmap contains one bit for each packet of a segment. This bitmap represents retransmission requests for those packets where there is a "0" in the corresponding position of the bitmap.

2.3 Rate control operations

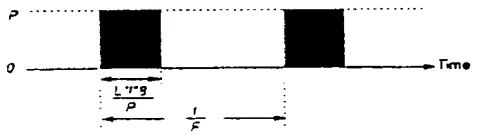


Figure 3: Bursty Rate Specification

There are several rate specification schemes proposed in the literature, e.g, an average rate calculated on a given time interval [18], a peak rate, an average rate, and a burst factor [1], or a leaky-bucket scheme as proposed for ATM. We choose a burst-rate specification for its simplicity. Let l be the packet size in bytes and L be the burst length in packets. Assume that the peak rate of transmission is P bits/s and the desired burst rate is F bursts/s. Then, the rate specification R consists of three parameters as follows:

$$R = (P, F, L) \quad where\ F \leq \frac{P}{L \times l \times 8}$$

The average rate in correspondence with the bursty rate R is $A = F \times L \times l \times 8$. Figure 3 depicts this bursty rate specification in time domain. A burst in this specification will usually be mapped to a segment of the applications.

Let BT be the burst timer with period $1/F$ and TC represent the transmission counter for a burst. Assume that there is a transmission queue XMTQ holding all the packets eligible for transmission. The burst transmission service routine will operate as follows when called upon BT timeout:

1. reset timer $BT = 1/F$.
2. set $TC = 0$.
3. if $TC < L$ and XMTQ not empty,
   remove and transmit the first packet from the queue and increment TC;
   else if $TC < L$ and XMTQ is empty,
   transmit a keep-alive packet or an end-of-stream packet and then return;
   otherwise return.
4. go back to step 3.

2.4 Window control operations

Receiver side operations

The receiver window-control functions determine the position of the window in the segment sequence space and the size of the window. Window status needs to be updated only when segments are successfully accepted at the receiver.

Let R_l be the lowest segment number within the receiving window, R_h the highest segment number within the receiving window, and W the current window size. Only packets with segment numbers in the range $[R_l, R_h]$ can be accepted by the receiver. There are two main steps in updating the window:

1. The receiver may adjust the window size W if there is a long-term change of processing speed in receiving application.
2. If for some segment i ($R_l \leq i \leq R_h$), all segments j ($R_l \leq j \leq i$) are successfully accepted by the application, the window is advanced by setting $R_l = i + 1$ and $R_h = \max(R_h, R_l + W - 1)$; a window advancement message will be sent to the sender with or without delay at the receiver's discretion.

Sender side operations

The window control function at the sender keeps track of the receiver's window advancement messages and updates the window at the sender as necessary. Let S_l and S_h specify the lower and upper edge of the sending window respectively. Only packets of the segments with segment numbers in the range $[S_l, S_h]$ are eligible for transmission by the rate-controlled transmission mechanism.

When a PACK or SNAK packet correctly arrives at the sender, its ACK segment NO and window upper limit fields contain respectively the R_l and R_h values advertised by the receiver. The sender updates the sending window as follows:

1. set $S_l = \max(S_l,$ ACK_segment_NO).
2. set $S_h = \max(S_h,$ window_upper_limit).

3 Performance study

Given the design of the two-level flow control scheme, two important performance questions need to be examined. (1) How does the window control perform given that the rate is guaranteed? (2) Is it necessary for the size of the receiver's buffer to be as large

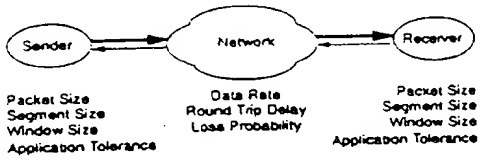


Figure 6: Simulation Configuration

Discrete event simulation

Simulation is used to study the effect of significant packet losses and the consequence of advertising windows larger than buffer sizes. The simulation configuration is shown in Figure 6.

The underlying network is modeled as a "black box" characterized by a connection data bandwidth, a round trip delay, and a packet loss process. The end system consists of two transport entities, a sender and a receiver. The end system parameters include packet size, segment size, window size, and application error tolerance. Performance measures of main interest are throughput efficiency, average end-to-end delay, and maximum end-to-end delay of a segment.

There are other parameters that can affect the performance of an error control scheme. For example, the total number of data segments to be transported and the amount of physical memory in the sender and receiver. We minimized these effects in the simulation by transmitting a large amount of data ($\geq 10^7$ packets) and by assuming very fast processors with large memories.

The following specific assumptions are made for the simulation:

- The round trip delay (RTD) on the connection does not vary in the duration of data transfer and the delay on each direction is RTD/2.
- All control packets are sent "out-of-band" which means they do not consume the bandwidth of data connection.
- An acknowledgment (PACK or SNAK) always comes back to the sender t_a time after the transmission of the last packet of a packet group, where $t_a = RTD + 3 \times t_p$ (same as that for the analyses).

3.1 Numerical results

This section presents and discusses a selected set of results from the simulation study. More detailed results have been reported in [11].

Window size requirement

Figure 7 shows throughput efficiency against window size. The results from both the analysis and the simulation are shown for comparison. The connected lines correspond to the analysis and the discrete symbols are for simulation. The loss probability was varied to obtain a family of three data sets which are labeled with different symbols in the figure. The error control was providing 100% loss recovery. The following can be observed from this plot:

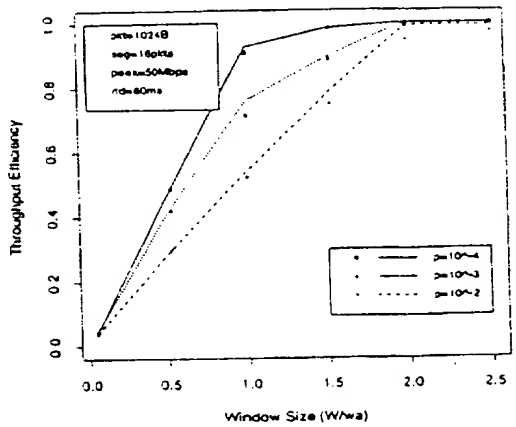


Figure 7: Throughput Efficiency – Analysis vs. Simulation

- With very small loss probability ($\leq 10^{-4}$), a window size of $(1 + w_a)$ is sufficient for achieving close to maximum throughput efficiency. Because with no loss, a window of size $(1 + w_a)$ is sufficient to keep the sender busy all the time.
- Larger window sizes are necessary to achieve the same efficiency when loss probability is higher. However, with loss probability of up to 10^{-2}, a window size of about 2.5 times bandwidth-delay product can achieve almost perfect throughput efficiency. The factor of 2.5 is in the same range as that found by Doshi et al. in SNR study.
- The throughput efficiency expressions derived in Section 3 seem to provide quite accurate prediction for loss probabilities up to 10^{-3}. But as expected, the analysis consistently predicts higher throughput due to the optimistic assumption about packet loss.

Buffer requirement

In an end-to-end communication, the sending buffer size always has to be at least as large as the window size because there can only be as many outstanding segments as there are sending segment buffers. The receiving buffer, on the other hand, can have a different size than the end-to-end window size. In particular, it is desirable to use a smaller receiving buffer if possible due to the constraint of physical memory sizes.

Indeed, under error-free conditions, the receiving buffer requirement is usually much smaller than the window size. Assume that the application reserves a buffer area for communication protocols to directly put new data in and for itself to do the processing. Then, two segment slots are sufficient for continuous receiving and processing operation. However, when there are errors or when the speeds of the protocol

Throughput efficiency analysis

Full analysis of error control schemes using selective ACK is very difficult. Existing studies (e.g., [14, 19]) have made simplifying assumptions such as absence of window flow control and no overlap between transmission and retransmission. These assumptions, though still allowing one to demonstrate the superiority of the selective ACK over the cumulative ACK, are too unrealistic for the operations of real schemes such as the one proposed. The approach used by Doshi and his colleagues in the analysis of SNR protocol is the only exception to this account (for the protocol see [15] and for analysis [7]), and it is the main inspiration for the throughput analysis to be presented.

In addition to those assumptions made for the delay analysis, we assume that the probability of requiring more than one retransmission for each packet, i.e., $(1-(1-p^2)^s)$, is negligible.

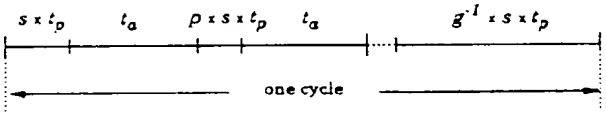


Figure 5: A Regenerative Cycle

Let $g = [1-(1-p)^s]$ denote the probability that a segment will require at least one packet retransmission. Activities on the connection follow regenerative cycles shown in Figure 5. Each time period, from left to right, as marked by vertical bars in the cycle is explained as follows:

1. The beginning of a cycle is marked by the transmission of a new segment which will have some packets lost.
2. After t_a time, a retransmission request (i.e., SNAK) comes back and $(p \times s)$ lost packets are retransmitted on the average.
3. According to the assumptions, all retransmissions will be successful and a positive acknowledgment (ACK) will be received in another t_a time; Depending on the actual size of the window, the two corresponding time periods (labeled as t_a in the figure) for acknowledgment may be utilized for transmitting new segments or left idle.
4. If more new segments are transmitted during the waiting periods for acknowledgment, some of the lost packets may have to be retransmitted in the period marked by the dotted line.
5. After recovery from the earlier losses, approximately g^{-1} segments can be delivered without any loss, until the next packet loss which will start the next cycle.

Once the regenerative cycle is identified, throughput efficiency can equivalently be defined as the ratio of the actual number of segments delivered in one cycle, to the ideal number of segments that could be delivered if there were no losses and the window size were infinite.

Let W be the number of segments in a window and $w_a = t_a/t_s$. Also, let ThE be the throughput efficiency. An expression for ThE with a window size W such that $1 < W \leq (w_a + 1)$ is derived first. In this case, the actual number of segments delivered in one cycle is:

$$N_{actual} \approx g^{-1} + W$$

because the window limits the number of segment transmissions to W during the whole time period $(2t_a + t_s + pt_s)$ and there are g^{-1} segments that can be delivered without loss. Now consider how many segments can be delivered under ideal condition. First of all, there can be $(2w_a + 1 + p)$ segments delivered during the time $(2t_a + t_s + pt_s)$. Since the smaller window forces the sender to wait from time to time during the transmission of the g^{-1} loss-free segments, it equivalently requires $\frac{(w_a+1)}{W}t_s$ time to deliver each segment. Thus, the total time for g^{-1} segments is $\frac{(w_a+1)}{W}g^{-1}t_s$. Ideally, exactly $\frac{(w_a+1)}{W}g^{-1}$ segments could have been delivered. The ideal number of segments deliverable is therefore:

$$N_{ideal} \approx 2w_a + 1 + p + \frac{(w_a+1)}{W}g^{-1}$$

and by definition:

$$\begin{aligned} ThE &\approx \frac{N_{actual}}{N_{ideal}} \\ &\approx \frac{g^{-1} + W}{2w_a + 1 + p + \frac{(w_a+1)}{W}g^{-1}} \end{aligned}$$

Similarly, the following expression can be derived for $(w_a + 1) < W < (2w_a + 1)$:

$$ThE \approx \frac{w_a + \min(W - w_a, w_a(1-p)) + g^{-1}}{2w_a + 1 + p + \min(W - w_a, w_a(1-p)) + g^{-1}}$$

Again, the numerator represents the actual number of segments delivered with the scheme in one cycle and the denominator is the number of segments deliverable under the ideal condition. Note that due to the simplifying assumptions made, the identified cycle is only appropriate for deriving expressions with $1 < W < (2w_a + 1)$. For $W = 1$, using the average end-to-end delay D_s given earlier:

$$ThE \approx \frac{t_s}{D_s}$$

When $W \geq 2w_a + 1$, the only reduction in throughput is due to retransmission, therefore:

$$ThE \approx (1-p)$$

Numerical results obtained using these expressions will be shown in Section 3.1.

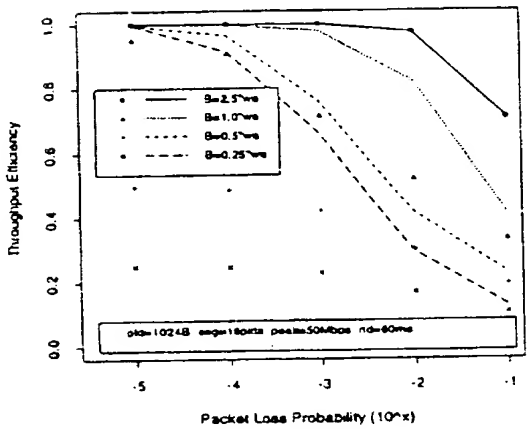


Figure 10: Advertising Larger Windows

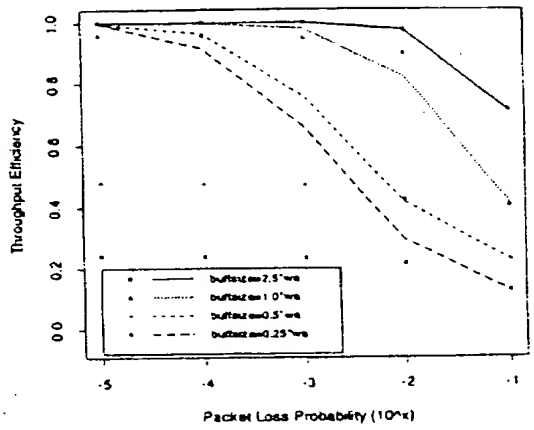


Figure 11: Hard ACK vs. Soft ACK - Throughput

are plotted in different line styles, while discrete points represent the results achieved when window sizes were kept the same as the receiving buffer sizes. In the case of advertising larger windows, the strategy used is hard ACK and the window is fixed at $2.5 \times w_a$. The vertical axis represents the throughput efficiency and the horizontal axis is the logarithmic value of random packet loss probability. Clearly, by advertising windows larger than the actual buffer size, significantly higher throughput is achieved. This holds for the whole range of loss probabilities studied and for buffer sizes as small as 1/4 of the bandwidth-delay product. In particular, over 70% higher throughput is achieved by advertising the large window for a buffer size of $0.25 \times w_a$ when loss probability is 10^{-5}.

The key advantage of advertising larger windows is that it allows the sender to continue transmitting data to fully utilize the connection. Although a segment may be dropped when it finds no buffer space when arriving at the receiver, such an event is not expected to occur very often because data will normally arrive successfully in sequence at the receiver and get consumed immediately, thus freeing more buffer space.

3.1.3 Hard ACK versus soft ACK performance

Plots in Figure 11 and 12 show comparisons between the hard ACK strategy and the soft ACK strategy. Figure 11 compares the throughput performance and Figure 12 shows the maximum segment delay. At all times, the end-to-end window size is fixed at $2.5 \times w_a$ and only buffer sizes are varied. The results from the hard ACK strategy are plotted using lines and the soft ACK results are shown as discrete symbols.

From Figure 11 it is seen that when the buffer size is equal to the window size $2.5 \times w_a$, the two strategies perform exactly the same (the solid line and the circles match) because there is no discard of segments at the receiver. But as the actual buffer size decreases, performance of the soft ACK strategy deteriorates dramatically. The most significant drop in throughput occurs when the buffer size is reduced to $0.5 \times w_a$. The performance difference between the two strategies remains significant until the loss probability approaches 10^{-1}, by that time the hard ACK performs as bad as the soft ACK due to the tremendous loss.

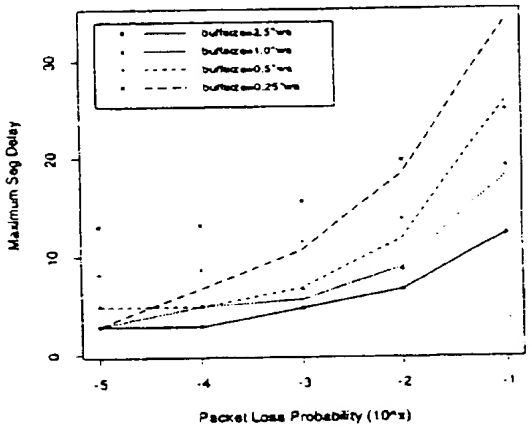


Figure 12: Hard ACK vs. Soft ACK - Maximum Delay

With the maximum segment delay (Figure 12), the hard ACK-strategy also outperforms the soft ACK strategy for loss probabilities up to 10^{-2}. Again, th two strategies perform identically when the receivin buffer is as large as the end-to-end window, just a expected. However, the soft ACK strategy seems t give lower maximum delay when the loss probabilit is very high, as seen for the loss probability of 10^{-1} i the plot. It should be noted that with the loss proba bility of 10^{-5}, the plot shows a higher maximum dela for buffer size $0.5 \times w_a$ than for buffer size $0.25 \times w_a$ This is believed to be an artifact of insufficient simu lation time for this extremely low loss case.

and the application do not match perfectly, data segments may have to be dropped or overwritten if no additional buffer slots are provided. An interesting question is: how effective is it to advertise larger windows for achieving higher performance? Two different acknowledgment strategies can be used for operating with over-advertised windows.

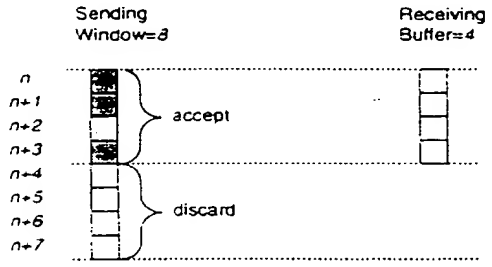


Figure 8: Hard ACK Strategy

3.1.1 Hard ACK and soft ACK strategies

Hard ACK and soft ACK are two types of acknowledgment strategies that can be used when the receiver is advertising a window larger than the actual receiving buffer available. These strategies have been considered for TCP extensions but were only briefly discussed on the end-to-end communications mailing list recently. To this date, there has been no performance study done to evaluate the effectiveness of these strategies, to the best of the author's knowledge. We address this need through simulation studies. The hard ACK and soft ACK strategies will be defined first and all results should be interpreted with respect to these definitions.

Let $(n, n+1, \ldots n+W-1)$ be the current window of size W that the receiver has advertised to the sender. Let B be the actual buffer size and $B < W$. Figure 8 illustrates how *hard* ACK strategy works with $W = 8$ and $B = 4$. It is seen that the receiver simply discards any segments with segment numbers outside the range $[n..n+3]$. All segments accepted can be acknowledged with certainty.

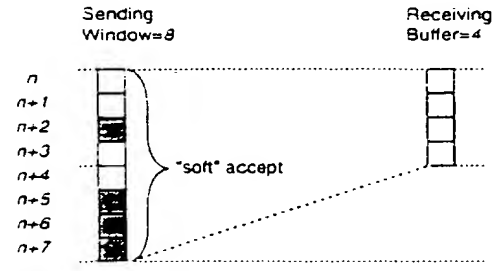


Figure 9: Soft ACK Strategy

The *soft* ACK strategy, however, will save a segment into the receiving buffer as long as it is inside the current window $[n, n+1, \ldots n+7]$, and there is space available for the segment, as shown in Figure 9. Because of retransmission and resequencing inside networks, a scenario like the one shown in the figure can occur. In this case, segments $n+2$, $n+5$, $n+6$, and $n+7$ arrived at the receiver before segments n, $n+1$, and $n+3$, and filled up the buffer. In order to deliver the next contiguous block of data to the application, some of the saved segments (e.g., $n+5$, $n+6$, and $n+7$) will have to be discarded to make room for segments n, $n+1$, and $n+3$. This means that when those out-of-order segments are received, they could only be acknowledged with a special message indicating their successful arrival at the receiver but the sender should not release those segments until further acknowledgment is received. The name soft ACK reflects exactly this special requirement.

It is clear that the soft ACK strategy requires proper specification of a *replacement policy* in order to be used. Specifically, when a segment needs to be discarded from the receiving buffer and there are multiple segments to choose from, how does one decide which segment to discard? Recall that the receiving buffer contains multiple segments which consist of many packets each. The following soft ACK algorithm is used in this study:

Upon successfully receiving a packet with segment number N_{in}:

1. If the packet belongs to a segment already in the buffer, accept it.
2. If the packet is new and there is free segment slot, accept the packet and reserve a slot.
3. If the packet is new and the buffer is all reserved, find the segment with the largest sequence number N_{old} in the buffer; if $N_{in} < N_{old}$ then discard segment N_{old} from the buffer and save the new packet in its slot; otherwise, discard the new packet.

Three simulation experiments have been conducted to achieve two goals: (1) to quantify the advantage of advertising larger windows and (2) to compare quantitatively the hard ACK and soft ACK strategies.

In the first experiment, the end-to-end window was set to be the same size as the receiving buffer all the time. The system was simulated for buffer sizes of 0.25, 0.5, 1.0, and 2.5 times the bandwidth-delay product w_a. For the second experiment, the simulator implemented the hard ACK strategy and the same set of buffer sizes were simulated, except that the end-to-end window was fixed at $2.5 \times w_a$. The third experiment repeated the set of conditions for experiment 2 but with the soft ACK strategy. In all three experiments, reliable simulation results were obtained by simulating 10 million packets. The results of these experiments are summarized in the next few paragraphs.

3.1.2 Larger window advantage

Figure 10 compares the throughput efficiencies achieved when the end-to-end window size is the same as the receiving buffer size and when larger windows are used. The results from advertising larger windows

[2] Cheriton, David, "VMTP: A Transport Protocol for the Next Generation of Computer Systems", *SIGCOMM '86 Symposium: Communications Architectures and Protocols (Computer Communication Review)*, Vol. 16, No. 3, ACM, New York, 1986, pp. 406–415.

[3] Chesson, Greg, et al., "XTP Protocol Definition", Revision 3.1, Protocol Engines, Inc., PEI 88–13, Santa Barbara, Calif., 1988.

[4] Clark, David D., Mark L. Lambert, and LiXia Zhang, "NETBLT: A High Throughput Transport Protocol", *SIGCOMM '87 Symposium: Frontiers in Computer Communications Technology (Computer Communication Review)*, Vol. 17, No. 5, ACM, New York, 1987, pp. 353–359.

[5] Cranor, Charles D. and Gurudatta M. Parulkar, "An Implementation Model for Connection-oriented Internet Protocols," *Journal of Internetworking: Research and Experience*, 12/93.

[6] Doeringer, Willibald A., et al., "A Survey of Light-Weight Transport Protocols for High-Speed Networks", *IEEE Trans. Communications*, Vol. 38, No. 11, November 1990, pp. 2025–2039.

[7] Doshi, B. T., et al., "Error and Flow Control Performance of a High Speed Protocol", internal draft, AT&T Bell Laboratories, 1991.

[8] Department of Defense, "Transmission Control Protocol", *MIL-STD-1778*, 20 May 1983.

[9] Gong, Fengmin and Gurudatta M. Parulkar, "Segment Streaming for Efficient Pipelined Televisualization", *Conference Record, IEEE Military Communications Conference*, Vol. 3, October 11–14, 1992, San Diego, pp. 991–997.

[10] Gong, Fengmin and Gurudatta M. Parulkar, "Error Control in High-Speed Networking Environments", to appear in Technical Proceedings of the 1994 Tactical Communications Conference, Fort Wayne, Indiana, May 10-12, 1994.

[11] Gong, Fengmin, *A Transport Solution for Pipelined Network Computing*, D.Sc. dissertation, Washington University Computer Science Department, St. Louis, December 1992.

[12] Jain, Raj, "Congestion Control in Computer Networks: Issues and Trends", *IEEE Network Magazine*, May 1990, pp. 24–30.

[13] Mazraani, Tony Y. and Gurudatta M. Parulkar, "Specification of a Multipoint Congram-Oriented High Performance Internet Protocol", *Proceedings of the Ninth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'90)* IEEE Computer Society, Washington D.C., June 1990.

[14] Mukherjee, Amarnath, *Analysis of Error Control and Congestion Control Protocols*, Ph.D. dissertation, Department of Computer Science, University of Wisconsin, November 1990.

[15] Netravali, Arun N., W. D. Roome, and K. Sabnani, "Design and Implementation of a High-Speed Transport Protocol", *IEEE Trans. on Communications*, Vol. 38, No. 11, November 1990, pp. 2010–2024.

[16] Papadopoulos, Christos and Gurudatta M. Parulkar, "Experimental Evaluation of SunOS IPC and TCP/IP Protocol Implementation" IEEE/ACM Transactions on Networking, Vol. 1 number 2, April 1993, pp. 199–216

[17] Sterbenz, James P. G., Gurudatta M. Parulkar, "Axon: Network Virtual Storage Design", *Computer Communication Review*, Vol. 20, No. 2, ACM SIGCOMM, New York, April 1990, pp. 50–56.

[18] Zhang, LiXia, "VirtualClock: A New Traffic Control Algorithm for Packet Switching Networks", *Proc. ACM SIGCOMM '90*, Philadelphia, PA, Sept. 1990, pp. 19–29.

[19] Zhou, Xiao You and Ahmed E. Kamal, "Automatic Repeat-Request Protocols and Their Queueing Analysis", *Computer Communications*, Vol. 13, No. 5, June 1990, pp. 298–311.

Why is the soft ACK strategy performing worse than the hard ACK? A careful examination of the soft ACK operation revealed the following behavior: When the window size is larger than the receiving buffer size, a packet loss or corruption in the network can lead the soft ACK into a state in which, periodically a number of segments will be first saved into the buffer but only to be replaced (thus discarded) later by segments with smaller sequence numbers. This state will persist even if there is no more loss in the network.

Note that this behavior not only causes spurious retransmissions, but also delays retransmission requests. Therefore, it leads to lower throughput efficiency and longer delay. This abnormal behavior also explains why the soft ACK performance is poor even with very low loss probability.

Intuitively, if the buffer size is very close to the window size, the soft ACK could possibly offer a better performance than the hard ACK strategy. But this margin seems to be so narrow that it is not visible within the range of parameters simulated. It is possible that by using a more elaborate saving and replacing policy, the abnormal behavior can be avoided and a better performance may be achieved using the soft ACK strategy. However, it is doubtful that more complex policies will allow the soft ACK to achieve much better performance than the hard ACK strategy.

4 Implementation

The proposed flow control scheme has been implemented along with an error control scheme as part of a segment streaming transport protocol (SSTP) inside the SunOS 4.0.3 kernel. SSTP is built on top of a connection-oriented internet protocol (COIP) [5, 13].

Extensive trace data has been collected that verified the flow control function. We also measured the throughput performance of the SSTP implementation using both custom software and a kernel probe technique developed by Papadopoulos [16]. The protocol processing delay results are summarized in Table 1. Protocol processing delays at both the sending and receiving ends are measured on a large number of packets. The resulting average is shown in the second column of the table. Note that this delay measure does not include the time for copying data from the application space to the kernel space or vice versa. In the last column, we also show the theoretical throughput corresponding to the given processing delay.

Send/Recv	Avg. Delay	Throughput
Send	273 μs	30 Mbps
Recv	310 μs	26 Mbps

Table 1: Per Packet Processing Delay

It is worth noting that the corresponding theoretical throughput for TCP/IP has been found to be about 22 Mbps. Thus, SSTP/COIP is about 20% faster than existing TCP/IP implementation. While TCP/IP implementation has been carefully crafted over the years, very little effort has been made to optimize implementations of SSTP or COIP. Furthermore, efficiency of the SSTP/COIP implementations are limited by the mbuf-based memory management scheme and by the lack of efficient timer support in the operating system kernel. We expect significant performance improvement for SSTP/COIP with the removal of these constraints and with additional hardware assistance. Since the control mechanisms of SSTP have been designed to maintain efficient operation even in large bandwidth-delay product networks, we expect SSTP to perform much better than TCP in high-speed network environments. More detailed results of the implementation are reported in [11].

5 Summary

This paper has presented a two-level flow control scheme that can provide efficient support for new distributed scientific computing applications that exchange a stream of data segments in high speed networks. The rate control ensures that the data sources (application processes) do not use more bandwidth than requested at the connection setup. This should help the underlying network to perform effective resource management and provide guaranteed services for applications. The window control provides end-to-end speed matching at the segment level and it is decoupled from the congestion control complication. The window advancement message is also separated from the error control acknowledgment to avoid interference with one another.

The paper also explored in detail a number of performance questions about the two-level scheme. The size requirements of the end-to-end window as well as the receiving buffer have been studied with analysis and extensive simulations. The results suggest: (1) with the two-level scheme an end-to-end window size of 2.5 times the bandwidth-delay product is sufficient for achieving nearly ideal throughput efficiency; (2) the receiving buffer size does not need to match the window size to attain very high performance; (3) the hard ACK strategy is far superior to the soft ACK strategy not only in performance, but also in the implementation complexity. Our measurement results show that even the primitive SSTP/COIP implementation performs significantly better than the well-crafted TCP/IP protocol. However, it should be pointed out that the proposed flow control scheme is intended for streaming of large set of data segments. Its comparison with TCP should be interpreted in this application domain.

Acknowledgment

Most of this work has been done at Washington University and it was supported in part by the National Science Foundation, and an industrial consortium of Ascom Timeplex, Bellcore, BNR, DEC, Italtel SIT, NEC, NTT, and SynOptics. We thank the anonymous reviewers for their very helpful comments.

References

[1] Akhtar, Shahid, *Congestion Control in a Fast Packet Switching Network*, Wash. U. CS Dept., M.S. thesis, St. Louis, Dec. 1987.